INTERNATIONAL TABLES
FOR
CRYSTALLOGRAPHY

Volume F

CRYSTALLOGRAPHY OF BIOLOGICAL MACROMOLECULES

Edited by

E. ARNOLD, D. M. HIMMEL AND M. G. ROSSMANN

Second Edition

Published for
THE INTERNATIONAL UNION OF CRYSTALLOGRAPHY
by
John Wiley & Sons, Ltd

2011

Dedicated to
DAVID C. PHILLIPS

The editors wish to express their special thanks to David Phillips (Lord Phillips of Ellesmere) and Professor Louise Johnson for contributing an exceptional chapter on the structure determination of hen egg-white lysozyme. Although Chapter 25.1 describes the first structural investigations of an enzyme, the procedures used are still as fresh and important today as they were 45 years ago, and this chapter is strongly recommended to students of both crystallography and enzymology. Completion of this chapter was David's last scientific accomplishment only a few weeks before his death. This volume of *International Tables for Crystallography* is dedicated to the memory of David C. Phillips in recognition of his pivotal contributions to the foundations of the crystallography of biological macromolecules.

Advisors and Advisory Board

Advisors: J. Drenth, A Liljas. **Advisory Board:** U. W. Arndt, E. N. Baker, H. M. Berman, T. L. Blundell, M. Bolognesi, A. T. Brunger, C. E. Bugg, R. Chandrasekaran, P. M. Colman, D. R. Davies, J. Deisenhofer, R. E. Dickerson, G. G. Dodson, H. Eklund, R. Giegé, J. P. Glusker, S. C. Harrison, W. G. J. Hol, K. C. Holmes, L. N. Johnson, K. K. Kannan, S.-H. Kim, A. Klug, D. Moras, R. J. Read, T. J. Richmond, G. E. Schulz, P. B. Sigler,† D. I. Stuart, T. Tsukihara, M. Vijayan, A. Yonath.

Contributing authors

P. D. Adams: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 18.6, 18.11]

P. V. Afonine: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

F. H. Allen: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

W. B. Arendall: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

U. W. Arndt: Laboratory of Molecular Biology, Medical Research Council, Hills Road, Cambridge CB2 2QH, England.† [6.1]

E. Arnold: Biomolecular Crystallography Laboratory, Center for Advanced Biotechnology and Medicine & Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA. [1.1, 1.4, 13.4]

P. Auffinger: Institut de Biologie Moléculaire et Cellulaire, Centre National de la Recherche Scientifique, Unité Propre de Recherche 9002, Université de Strasbourg, 15 Rue René Descartes, F-67084 Strasbourg, France. [23.6]

E. N. Baker: School of Biological Sciences, University of Auckland, Private Bag 92-109, Auckland, New Zealand. [22.3]

T. S. Baker: Departments of Chemistry & Biochemistry and Molecular Biology, University of California, San Diego, 9500 Gilman Drive MC-0378, La Jolla, CA 92093-0378, USA. [19.6]

W. Baxter: Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA. [19.8]

C. Bebeacua: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

C. G. van Beek: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA.‡ [11.5]

J. A. Bell: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

J. Berendzen: Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.3]

H. M. Berman: Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA. [21.2, 24.1, 24.2]

C. C. F. Blake: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

D. M. Blow: Biophysics Group, Blackett Laboratory, Imperial College of Science, Technology & Medicine, London SW7 2BW, England.† [13.1]

T. L. Blundell: Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge, CB2 1GA, England. [12.1]

R. Bolotovsky: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA.¶ [11.5]

D. Borek: UT Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA. [11.4]

G. Bricogne: Laboratory of Molecular Biology, Medical Research Council, Cambridge CB2 2QH, England. [16.2]

A. T. Brunger: The Howard Hughes Medical Institute, and Departments of Molecular and Cellular Physiology, Neurology and Neurological Sciences, and Stanford Synchrotron Radiation Laboratory, Stanford Universty, 1201 Welch Road, MSLS P210, Stanford, CA 94305-5489, USA. [18.2, 18.6]

G. Bunkóczi: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

Y. Cao: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

H. L. Carrell: The Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, PA 19111, USA. [5.1]

D. Carvin: Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Field, London WC2A 3PX, England. [12.1]

R. Chandrasekaran: Whistler Center for Carbohydrate Research, Purdue University, West Lafayette, IN 47907, USA. [19.5]

M. S. Chapman: Department of Chemistry & Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306-4380, USA. [22.2]

L. Chen: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

V. B. Chen: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

W. Chiu: Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030, USA. [19.2]

K. H. Choi: Department of Biochemistry and Molecular Biology, 6.614C Basic Science, The University of Texas Medical Branch, University Blvd, Galveston, TX 77555-0647, USA. [4.4]

J. C. Cole: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

M. L. Connolly: 1259 El Camino Real 184, Menlo Park, CA 94025, USA. [22.2]

C. Cork: Structural Proteomics Development Group, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [9.2]

K. Cowtan: Department of Chemistry, University of York, York YO10 5DD, UK. [15.1, 15.3, 17.1]

D. W. J. Cruickshank: Chemistry Department, UMIST, Manchester M60 1QD, England.† [18.5]

J. de la Cruz: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

M. Cymborowski: Department of Molecular Physiology and Biological Physics, University of Virginia, 1300 Jefferson Park Avenue, Charlottesville, VA 22908, USA. [11.4]

V. M. Dadarlat: Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana 47907-1333, USA. [20.2]

U. Das: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

Z. Dauter: NCI Frederick & Argonne National Laboratory, Building 202, Argonne, IL 60439, USA. [9.1, 18.4]

I. W. Davis: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA.†† [18.11]

T. Day: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

W. L. DeLano: Graduate Group in Biophysics, Box 0448, University of California, San Francisco, CA 94143, USA. [18.6]

Z. S. Derewenda: Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, VA 22908-0736, USA. [4.3]

---

† Deceased.

‡ Present address: RJ Lee Instruments, 515 Pleasant Valley Road, Trafford, PA 15085, USA.

§ Present address: Kent House, 19 The Warren, Cromer, Norfolk NR27 0AR, England.

¶ Present address: Philips Analytical Inc., 12 Michigan Drive, Natick, MA 01760, USA.

† Deceased.

†† Present address: GrassRoots Biotechnology, 598 Airport Boulevard, Morrisville, NC 27560, USA.

CONTRIBUTING AUTHORS

R. E. Dickerson: Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90095-1570, USA. [23.4]

J. Drenth: Laboratory of Biophysical Chemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands. [2.1]

S. Dutta: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

O. Dym: UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, UCLA, Box 951570, Los Angeles, CA 90095-1570, USA. [21.3]

T. Earnest: Structural Proteomics Development Group, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [9.2]

N. Echols: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

E. F. Eikenberry: Swiss Light Source, Paul Scherrer Institut, 5232 Villigen PSI, Switzerland. [7.1, 7.2]

H. M. Einspahr: PO Box 6483, Lawrenceville, NJ 08648-0483, USA. [2.2]

D. Eisenberg: UCLA–DOE Laboratory of Structural Biology and Molecular Medicine, Department of Chemistry & Biochemistry, Molecular Biology Institute and Department of Biological Chemistry, UCLA, Los Angeles, CA 90095-1570, USA. [21.3]

P. Emsley: Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK. [17.1]

D. M. Engelman: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [19.4]

R. A. Engh: Pharmaceutical Research, Roche Diagnostics GmbH, Max Planck Institut für Biochemie, 82152 Martinsried, Germany. [18.3]

J. A. Ernst: Department of Protein Chemistry, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

R. Farid: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

Z. Feng: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

R. H. Fenn: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

J. Frank: Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, NY 10032, USA. [19.8]

D. T. Gallagher: Center for Advanced Research in Biotechnology of the Maryland Biotechnology Institute and National Institute of Standards and Technology, 9600 Gudelsky Dr., Rockville, MD 20850, USA. [24.3]

E. Gallicchio: BioMaPS Institute for Quantitative Biology, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [18.10]

E. F. Garman: Department of Biochemistry, South Parks Road, Oxford, OX1 3QU, England. [10.3]

M. Gerstein: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1]

R. Giegé: Machineries Traductionnelles, ARN, UPR 9002, IBMC du CNRS, 15 rue René Descartes, Strasbourg, 67084, France. [4.1]

C. J. Gilmore: Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, UK. [16.1]

J. P. Glusker: The Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, PA 19111, USA. [5.1]

T. Grant: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

P. Gros: Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. [18.6]

R. W. Grosse-Kunstleve: The Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.6, 18.11]

S. M. Gruner: Department of Physics, 162 Clark Hall, Cornell University, Ithaca, NY 14853-2501, USA. [7.1, 7.2]

J. R. Gunn: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

W. F. van Gunsteren: Laboratory of Physical Chemistry, ETH-Zentrum, 8092 Zürich, Switzerland. [20.1]

H. A. Hauptman: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

J. J. Headd: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.6]

M. van Heel: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

J. R. Helliwell: Department of Chemistry, University of Manchester, M13 9PL, England. [8.1]

R. Henderson: Medical Research Council, Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. [19.6]

W. A. Hendrickson: Department of Biochemistry, College of Physicians & Surgeons of Columbia University, 630 West 168th Street, New York, NY 10032, USA. [14.2]

K. Henrick: Protein Data Bank in Europe (PDBe), EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [24.1]

D. M. Himmel: Biomolecular Crystallography Laboratory, Center for Advanced Biotechnology and Medicine & Rutgers University, 679 Hoes Lane, Piscataway, NJ 08854-5638, USA. [1.1, 1.4]

P. S. Ho: Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, Colorado State University, Fort Collins, CO 80523-1870, USA. [23.6]

A. E. Hodel: Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322, USA.† [23.3]

W. G. J. Hol: Biomolecular Structure Center, Department of Biological Structure, Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195-7742, USA. [1.3]

L. Holm: EMBL–EBI, Cambridge CB10 1SD, England. [23.2]

H. Hope: Department of Chemistry, University of California, Davis, One Shields Ave, Davis, CA 95616-5295, USA. [10.1]

R. Huber: Max-Planck-Institut für Biochemie, 82152 Martinsried, Germany. [12.2, 18.3]

S. H. Hughes: National Cancer Institute, Frederick Cancer R&D Center, Frederick, MD 21702-1201, USA. [3.1]

L.-W. Hung: Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [18.11]

R. M. Immormino: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

S. A. Islam: Institute of Cancer Research, 44 Lincoln's Inn Fields, London WC2A 3PX, England. [12.1]

J.-S. Jiang: Protein Data Bank, Biology Department, Brookhaven National Laboratory, Upton, NY 11973-5000, USA. [18.6]

J. E. Johnson: Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037, USA. [19.3]

L. N. Johnson: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

W. Kabsch: Max-Planck-Institut für Medizinische Forschung, Abteilung Biophysik, Jahnstrasse 29, 69120 Heidelberg, Germany. [11.3, 11.6]

G. J. Kapral: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.6]

D. A. Keedy: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

G. Kleywegt: Protein Data Bank in Europe (PDBe), EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [21.1, 24.1]

R. Knott: Small Angle Scattering Facility, Australian Nuclear Science & Technology Organisation, Physics Division, PMB 1 Menai NSW 2234, Australia. [6.2]

D. F. Koenig: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.¶ [25.1]

A. A. Kossiakoff: Department of Biochemistry and Molecular Biology, CLSC 161A, University of Chicago, Chicago, IL 60637, USA. [19.1]

---

‡ Present address: 2 Second Avenue, Denvilles, Havant, Hampshire P09 2QP, England.

† Deceased.

§ Present address: Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 2QU, England.

¶ Present address unknown.

CONTRIBUTING AUTHORS

C. M. Koth: Department of Structural Biology, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

V. S. Lamzin: European Molecular Biology Laboratory (EMBL), Hamburg Outstation, c/o DESY, Notkestr. 85, 22603 Hamburg, Germany. [18.8]

R. A. Laskowski: Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England. [21.4]

A. Leith: Resource for Visualization of Biological Complexity, Wadsworth Center, Empire State Plaza, Albany, NY 12201-0509, USA. [19.8]

A. G. W. Leslie: MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England. [11.2]

R. Levy: BioMaPS Institute for Quantitative Biology, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA. [18.10]

C. Linnemayr: Division of Internal Medicine, Inflammation Research, University Hospital Zürich, Switzerland. [19.9]

B. Lohkamp: Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-171 77 Stockholm, Sweden. [17.1]

B. Lorber: UPR 9002, IBMC–CNRS, 15 rue René Descartes Cedex, Strasbourg, 67084, France. [4.1]

S. Ludtke: National Center for Macromolecular Imaging, Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [19.10]

V. Y. Lunin: Laboratory of Macromolecular Crystallography, Institute of Mathematical Problems of Biology of the Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russian Federation. [16.3]

M. W. MacArthur: Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, England. [21.4]

A. J. McCoy: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

A. McPherson: Department of Molecular Biology and Biochemistry, University of California, 560 Steinhaus, Irvine, CA 92697-3900, USA. [4.1]

P. Main: Department of Physics, University of York, York YO1 5DD, England. [15.1, 15.3]

G. A. Mair: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

J. Markley: BioMagResBank, University of Wisconsin–Madison, Madison, Wisconsin 53706, USA. [24.1]

B. W. Matthews: Institute of Molecular Biology, Howard Hughes Medical Institute and Department of Physics, University of Oregon, Eugene, OR 97403, USA. [14.1]

C. Mattos: Department of Molecular and Structural Biochemistry, North Carolina State University, 128 Polk Hall, Raleigh, NC 02795, USA. [23.5]

H. Michel: Max-Planck-Institut für Biophysik, Heinrich-Hoffmann-Strasse 7, D-60528 Frankfurt/Main, Germany. [4.2]

R. Miller: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

W. Minor: Department of Molecular Physiology and Biological Physics, University of Virginia, 1300 Jefferson Park Avenue, Charlottesville, VA 22908, USA. [11.4]

K. Moffat: Department of Biochemistry and Molecular Biology, The Center for Advanced Radiation Sources, and The Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois 60637, USA. [8.2]

P. B. Moore: Departments of Chemistry and Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [19.4]

N. W. Moriarty: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

L. W. Murray: Department of Biochemistry, Duke University, Durham, NC 27710, USA. [21.6]

G. N. Murshudov: York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, UK. [18.4]

H. Nakamura: PDBj, Institute for Protein Research, Osaka University, Japan. [24.1]

J. Navaza: Laboratoire de Génétique des Virus, CNRS-GIF, 1. Avenue de la Terrasse, 91198 Gif-sur-Yvette, France. [13.2]

A. C. T. North: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.§ [25.1]

R. Oeffner: Department of Haematology, University of Cambridge, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Cambridge CB2 0XY, England. [18.11]

J. W. H. Oldham: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.† [25.1]

A. J. Olson: The Scripps Research Institute, La Jolla, CA 92037, USA. [17.2]

C. Orengo: Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, England. [23.1]

Z. Otwinowski: UT Southwestern Medical Center at Dallas, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA. [11.4]

N. S. Pannu: Department of Mathematical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. [18.6]

S. Parkin: Department of Chemistry, University of Kentucky, Lexington, Kentucky, USA. [10.1]

A. Perrakis: European Molecular Biology Laboratory (EMBL), Grenoble Outstation, c/o ILL, Avenue des Martyrs, BP 156, 38042 Grenoble Cedex 9, France. [18.8]

D. C. Phillips: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.† [25.1]

A. Podjarny: Structural Biology, IGBMC, BP 163 Cedex, Illkirch, 67404, France. [16.3]

R. J. Poljak: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.¶ [25.1]

J. Pontius: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

R. Portugal: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England. [19.9]

C. B. Post: Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, Indiana 47907-1333, USA. [20.2]

F. A. Quiocho: Howard Hughes Medical Institute and Department of Biochemistry, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [23.3]

R. J. Read: Department of Haematology, University of Cambridge, CIMR, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 0XY, England. [15.2, 18.6, 18.11]

L. M. Rice: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA. [18.2, 18.6]

F. M. Richards: Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA. [22.1]

D. C. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

J. S. Richardson: Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA. [18.11, 21.5, 21.6]

J. Richelle: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium. [21.2]

D. Ringe: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, 415 South St, Waltham, MA 02254, USA. [23.5]

D. W. Rodgers: Department of Biochemistry, Chandler Medical Center, University of Kentucky, 800 Rose Street, Lexington, KY 40536-0298, USA. [10.2]

A. Rohou: Faculty of Natural Sciences, Division of Molecular Biosciences, Imperial College London, London SW7 2AZ, England.†† [19.9]

M. G. Rossmann: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [1.1, 1.2, 1.4, 11.1, 11.5, 13.4]

C. Sander: MIT Center for Genome Research, One Kendall Square, Cambridge, MA 02139, USA. [23.2]

---

‡ Present address unknown.

§ Present address: Prospect House, 27 Breary Lane, Bramhope, Leeds LS16 9AD, England.

† Deceased.

¶ Present address: CARB, 9600 Gudelsky Drive, Rockville, MD 20850, USA.

†† Present address: Rosenstiel Center, Brandeis University, 415 South Street, Waltham, MA 02454-9110, USA.

CONTRIBUTING AUTHORS

V. R. Sarma: Davy Faraday Research Laboratory, The Royal Institution, London W1X 4BS, England.‡ [25.1]

C. Sauter: Institut de Biologie Moléculaire et Cellulaire (IBMC), Centre National pour la Recherche Scientifique (CNRS), 15 rue René Descartes, Strasbourg, F-67084, France. [4.1]

M. R. Sawaya: Institute for Genomics and Proteomics, UCLA–DOE, 611 Young Drive East, Los Angeles, CA 90095, USA. [18.12]

M. Schatz: Image Science Software GmbH, Gillweg 3, D-14193 Berlin, Germany. [19.9]

R. Schmidt: Image Science Software GmbH, Gillweg 3, D-14193 Berlin, Germany. [19.9]

B. Schneider: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA.§ [24.2]

B. P. Schoenborn: Life Sciences Division M888, University of California, Los Alamos National Laboratory, Los Alamos, NM 8745, USA. [6.2]

D. Shapiro: Advanced Light Source, Lawrence Berkeley National Laboratory, 1 Cyclotron Rd, MS 2-400, Berkeley, CA 94720, USA. [9.3]

K. A. Sharp: E. R. Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104-6059, USA. [22.4]

G. M. Sheldrick: Lehrstuhl für Strukturchemie, Georg-August-Universität Göttingen, Tammannstrasse 4, D-37077 Göttingen, Germany. [16.1, 18.9]

T. Simonson: Laboratoire de Biologie Structurale (CNRS), IGBMC, 1 rue Laurent Fries, 67404 Illkirch (CU de Strasbourg), France. [18.6]

J. L. Smith: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [14.2]

J. C. H. Spence: Department of Physics, Arizona State University, Tempe, Arizona, 85287, USA. [1.4]

M. J. E. Sternberg: Institute of Cancer Research, 44 Lincoln's Inn Fields, London WC2A 3PX, England. [12.1]

A. M. Stock: Center for Advanced Biotechnology and Medicine, Howard Hughes Medical Institute and University of Medicine and Dentistry of New Jersey – Robert Wood Johnson Medical School, 679 Hoes Lane, Piscataway, NJ 08854-5627, USA. [3.1]

U. Stocker: Laboratory of Physical Chemistry, ETH-Zentrum, 8092 Zürich, Switzerland. [20.1]

G. Stubbs: Department of Molecular Biology, Vanderbilt University, Nashville, TN 37235, USA. [19.5]

M. T. Stubbs: Institut für Pharmazeutische Chemie der Philipps-Universität Marburg, Marbacher Weg 6, D-35032 Marburg, Germany. [12.2]

S. Sun: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907-1392, USA. [1.4]

M. W. Tate: Department of Physics, 162 Clark Hall, Cornell University, Ithaca, NY 14853-2501, USA. [7.1, 7.2]

L. F. Ten Eyck: San Diego Supercomputer Center 0505, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0505, USA. [18.1, 18.7]

A. Teplyakov: University of Maryland Biotechnology Institute, Rockville, MD 20850, USA.¶ [13.5]

T. C. Terwilliger: Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [14.3, 18.11]

J. Thornton: Biochemistry and Molecular Biology Department, University College London, Gower Street, London WC1E 6BT, England, and Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, England. [21.4, 23.1]

L. Tong: Department of Biological Sciences, Columbia University, New York, NY 10027, USA. [13.3]

D. E. Tronrud: Howard Hughes Medical Institute, Institute of Molecular Biology, 1229 University of Oregon, Eugene, OR 97403-1229, USA. [18.7]

Y. Tsai: Department of Chemistry and Biochemistry, University of California, Los Angeles, PO Box 951569, Los Angeles, CA 90095-1569, USA. [11.7]

H. Tsuruta: SSRL/SLAC & Department of Chemistry, Stanford University, PO Box 4349, MS69, Stanford, California 94309-0210, USA. [19.3]

M. Tung: Center for Advanced Research in Biotechnology of the Maryland Biotechnology Institute and National Institute of Standards and Technology, 9600 Gudelsky Dr., Rockville, MD 20850, USA. [24.3]

A. G. Urzhumtsev: Faculty of Sciences, University of Nancy 1, Vandoeuvre-lès-Nancy, 54506, France. [16.3]

I. Usón: Institució Catalana de Recerca i Estudis Avançats at IBMB–CSIC, Barcelona Science Park, Baldiri Reixach 15, 08028 Barcelona, Spain. [16.1]

A. Vagin: Structural Biology Laboratory, University of York, York YO10 5YW, England. [13.5, 21.2]

M. J. Vallejos: Department of Biochemistry and Molecular Biology, 1870 Campus Delivery, Colorado State University, Fort Collins, CO 80523-1870, USA. [23.6]

M. L. Verdonk: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England. [22.5]

C. L. M. J. Verlinde: Biomolecular Structure Center, Department of Biological Structure, Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195-7742, USA. [1.3]

C. A. Vernon: Department of Chemistry, University College London, Gower Street, London WC1E 6BT, England.† [25.1]

K. D. Watenpaugh: Structural, Analytical and Medicinal Chemistry, Pharmacia & Upjohn, Inc., Kalamazoo, MI 49001-0119, USA. [18.1]

C. M. Weeks: Hauptman–Woodward Medical Research Institute, Inc., 7000 Ellicott Street, Buffalo, NY 14203-1102, USA. [16.1]

M. S. Weiss: Helmholtz-Zentrum Berlin für Materialien und Energie, Macromolecular Crystallography (HZB-MX), Albert-Einstein-Str. 15, D-12489 Berlin, Germany. [2.2]

E. M. Westbrook: Molecular Biology Consortium, Argonne, Illinois 60439, USA. [5.2]

J. Westbrook: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

K. S. Wilson: York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England. [9.1, 18.4, 18.8]

S. J. Wodak: Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium, and EMBL–EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England. [21.2]

K. Wüthrich: Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg, CH-8093 Zürich, Switzerland. [19.7]

H. Yang: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

D. G. Yansura: Department of Antibody Engineering, Genentech, 1 DNA Way, South San Francisco, California 94080, USA. [3.2]

T. O. Yeates: Department of Chemistry and Biochemistry, University of California, Los Angeles, PO Box 951569, Los Angeles, CA 90095-1569, USA. [11.7, 18.12, 21.3]

J. Young: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

C. Zardecki: The Nucleic Acid Database, Department of Chemistry and Chemical Biology, Rutgers, The State University of New Jersey, Piscataway, New Jersey 08854, USA. [24.2]

K. Y. J. Zhang: Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N., Seattle, WA 90109, USA. [15.1, 15.3]

Z. Zhou: Schrödinger, 120 West 45th Street, 17th Floor, New York, NY 10036, USA. [18.10]

P. H. Zwart: Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [18.11]

---

‡ Present address: Department of Biochemistry, State University of New York at Stonybrook, Stonybrook, NY 11794-5215, USA.

§ Present address: Institute of Biotechnology, Academy of Science, Czech Republic.

¶ Present address: Centocor R&D Inc., Radnor, PA 19087, USA.

† Deceased.

Contents

PAGE

Preface (M. G. Rossmann and E. Arnold) xxviii

Preface to the second edition (E. Arnold, D. M. Himmel and M. G. Rossmann) xxix

PART 1. INTRODUCTION 1

1.1. Overview (E. Arnold, D. M. Himmel and M. G. Rossmann) 1

1.2. Historical background (M. G. Rossmann) 5

- **1.2.1. Introduction** 5
- **1.2.2. 1912 to the 1950s** 5
- **1.2.3. The first investigations of biological macromolecules** 6
- **1.2.4. Globular proteins in the 1950s** 7
- **1.2.5. The first protein structures (1957 to the 1970s)** 8
- **1.2.6. Technological developments (1958 to the 1980s)** 9
- **1.2.7. Meetings** 10

1.3. Macromolecular crystallography and medicine (W. G. J. Hol and C. L. M. J. Verlinde) 13

- **1.3.1. Introduction** 13
- **1.3.2. Crystallography and medicine** 13
- **1.3.3. Crystallography and genetic diseases** 14
- **1.3.4. Crystallography and development of novel pharmaceuticals** 15
- **1.3.5. Vaccines, immunology and crystallography** 26
- **1.3.6. Outlook and dreams** 26

1.4. Perspectives for the future (E. Arnold, M. G. Rossmann, D. M. Himmel, J. C. H. Spence and S. Sun) 39

- **1.4.1. Gazing into the crystal ball (E. Arnold)** 39
- **1.4.2. Brief comments on *Gazing into the crystal ball* (M. G. Rossmann)** 40
- **1.4.3. Additional comments on *Gazing into the crystal ball* (D. M. Himmel)** 41
- **1.4.4. *Gazing into the crystal ball* – the X-ray free-electron laser (J. C. H. Spence)** 41
- **1.4.5. Electron microscopy's impact on structural biology (S. Sun)** 42

PART 2. BASIC CRYSTALLOGRAPHY 45

2.1. Introduction to basic crystallography (J. Drenth) 45

- **2.1.1. Crystals** 45
- **2.1.2. Symmetry** 46
- **2.1.3. Point groups and crystal systems** 47
- **2.1.4. Basic diffraction physics** 52
- **2.1.5. Reciprocal space and the Ewald sphere** 57
- **2.1.6. Mosaicity and integrated reflection intensity** 58
- **2.1.7. Calculation of electron density** 60
- **2.1.8. Symmetry in the diffraction pattern** 60
- **2.1.9. The Patterson function** 62

2.2. Quality indicators in macromolecular crystallography: definitions and applications (H. M. Einspahr and M. S. Weiss) 64

- **2.2.1. Introduction** 64
- **2.2.2. Quality indicators for diffraction data** 65
- **2.2.3. Comparing different diffraction data sets** 68
- **2.2.4. Quality indicators for substructure determination** 68
- **2.2.5. Quality indicators for phase determination** 69
- **2.2.6. Quality indicators for density modification and phase improvement** 69

2.2.7. Quality indicators for molecular replacement .. 70
2.2.8. Quality indicators for refinement .. 71
2.2.9. Quality indicators for the refined model .. 71
2.2.10. Error estimation for the refined model .. 73
2.2.11. The most commonly used quality indicators .. 73

PART 3. TECHNIQUES OF MOLECULAR BIOLOGY .. 75

3.1. Preparing recombinant proteins for X-ray crystallography (S. H. Hughes and A. M. Stock) .. 75
3.1.1. Introduction .. 75
3.1.2. Overview .. 75
3.1.3. Engineering an expression construct .. 76
3.1.4. Expression systems .. 77
3.1.5. Protein purification .. 85
3.1.6. Characterization of the purified product .. 88
3.1.7. Reprise .. 89

3.2. Expression and purification of membrane proteins for structural studies (J. A. Ernst, D. G. Yansura and C. M. Koth) .. 92
3.2.1. Introduction .. 92
3.2.2. A consensus strategy for membrane-protein expression .. 92
3.2.3. A consensus strategy for membrane-protein purification .. 93
3.2.4. Common purification pitfalls and prioritized alternative strategies .. 96
3.2.5. Summary .. 96

PART 4. CRYSTALLIZATION .. 99

4.1. General methods (C. Sauter, B. Lorber, A. McPherson and R. Giegé) .. 99
4.1.1. Introduction .. 99
4.1.2. Main parameters that affect crystallization of macromolecules .. 100
4.1.3. Crystallization arrangements and classical methodologies .. 104
4.1.4. Advanced crystallization methodologies .. 107
4.1.5. From the macromolecule to perfect crystals: the physics view .. 111
4.1.6. How to crystallize a new macromolecule: the structural biology view .. 113
4.1.7. The future of protein crystal growth .. 115

4.2. Crystallization of membrane proteins (H. Michel) .. 122
4.2.1. Introduction .. 122
4.2.2. Principles of membrane-protein crystallization .. 122
4.2.3. General properties of detergents relevant to membrane-protein crystallization .. 123
4.2.4. The ‘small amphiphile concept’ .. 125
4.2.5. Membrane-protein crystallization with the help of antibody Fv fragments .. 126
4.2.6. Membrane-protein crystallization using cubic bicontinuous lipidic phases .. 126
4.2.7. General recommendations .. 126

4.3. Application of protein engineering to enhance crystallizability and improve crystal properties (Z. S. Derewenda) .. 129
4.3.1. Introduction .. 129
4.3.2. Microscopic aspects of protein crystallization .. 129
4.3.3. Engineering proteins with enhanced solubility .. 130
4.3.4. Optimization of target constructs .. 131
4.3.5. The use of fusion proteins for crystallization .. 131
4.3.6. Noncovalent crystallization chaperones .. 132
4.3.7. Removal of post-translational modifications .. 133
4.3.8. Stabilization of protein targets .. 134
4.3.9. Surface-entropy reduction (SER) .. 134

4.3.10. Improvement of crystal quality 135

4.3.11. Conclusions 135

4.4. High-throughput X-ray crystallography (K. H. Choi) 140

4.4.1. Introduction 140

4.4.2. Design of multiple constructs: bioinformatics analysis of genome sequences 140

4.4.3. Cloning 140

4.4.4. Protein expression and purification 142

4.4.5. Crystallization 142

4.4.6. Synchrotron data collection 143

PART 5. CRYSTAL PROPERTIES AND HANDLING 145

5.1. Crystal morphology, optical properties of crystals and crystal mounting (H. L. Carrell and J. P. Glusker) 145

5.1.1. Crystal morphology and optical properties 145

5.1.2. Crystal mounting 148

5.2. Crystal-density measurements (E. M. Westbrook) 152

5.2.1. Introduction 152

5.2.2. Solvent in macromolecular crystals 152

5.2.3. Matthews number 152

5.2.4. Algebraic concepts 152

5.2.5. Experimental estimation of hydration 153

5.2.6. Methods for measuring crystal density 153

5.2.7. How to handle the solvent density 156

PART 6. RADIATION SOURCES AND OPTICS 159

6.1. X-ray sources (U. W. Arndt) 159

6.1.1. Overview 159

6.1.2. Generation of X-rays 159

6.1.3. Properties of the X-ray beam 162

6.1.4. Beam conditioning 164

6.2. Neutron sources (B. P. Schoenborn and R. Knott) 168

6.2.1. Reactors 168

6.2.2. Spallation neutron sources 172

6.2.3. Summary 175

PART 7. X-RAY DETECTORS 177

7.1. Comparison of X-ray detectors (S. M. Gruner, E. F. Eikenberry and M. W. Tate) 177

7.1.1. Commonly used detectors: general considerations 177

7.1.2. Evaluating and comparing detectors 178

7.1.3. Characteristics of different detector approaches 179

7.1.4. Future detectors 181

7.2. CCD detectors (M. W. Tate, E. F. Eikenberry and S. M. Gruner) 183

7.2.1. Overview 183

7.2.2. CCD detector assembly 183

7.2.3. Calibration and correction 184

7.2.4. Detector system integration 186

7.2.5. Applications to macromolecular crystallography 187

7.2.6. Future of CCD detectors 187

PART 8. SYNCHROTRON CRYSTALLOGRAPHY .. 189

8.1. Synchrotron-radiation instrumentation, methods and scientific utilization (J. R. Helliwell) .. 189
- **8.1.1. Introduction** .. 189
- **8.1.2. The physics of SR** .. 189
- **8.1.3. Insertion devices (IDs)** .. 190
- **8.1.4. Beam characteristics delivered at the crystal sample** .. 191
- **8.1.5. Evolution of SR machines and experiments** .. 192
- **8.1.6. SR instrumentation** .. 194
- **8.1.7. SR monochromatic and Laue diffraction geometry** .. 195
- **8.1.8. Scientific utilization of SR in protein crystallography** .. 197
- **8.1.9. Concluding remarks** .. 200

8.2. Laue crystallography: time-resolved studies (K. Moffat) .. 205
- **8.2.1. Introduction** .. 205
- **8.2.2. Principles of Laue diffraction** .. 205
- **8.2.3. Practical considerations in the Laue technique** .. 206
- **8.2.4. The time-resolved experiment** .. 208
- **8.2.5. Conclusions** .. 209

PART 9. X-RAY DATA COLLECTION .. 211

9.1. Principles of monochromatic data collection (Z. Dauter and K. S. Wilson) .. 211
- **9.1.1. Introduction** .. 211
- **9.1.2. The components of a monochromatic X-ray experiment** .. 211
- **9.1.3. Data completeness** .. 211
- **9.1.4. X-ray sources** .. 211
- **9.1.5. Goniostat geometry** .. 212
- **9.1.6. Basis of the rotation method** .. 213
- **9.1.7. Rotation method: geometrical completeness** .. 217
- **9.1.8. Crystal-to-detector distance** .. 221
- **9.1.9. Wavelength** .. 222
- **9.1.10. Lysozyme as an example** .. 222
- **9.1.11. Rotation method: qualitative factors** .. 223
- **9.1.12. Radiation damage** .. 225
- **9.1.13. Relating data collection to the problem in hand** .. 226
- **9.1.14. The importance of low-resolution data** .. 228
- **9.1.15. Data quality over the whole resolution range** .. 228
- **9.1.16. Strategies for automated data acquisition** .. 229
- **9.1.17. Final remarks** .. 229

9.2. Robotic crystal loading (T. Earnest and C. Cork) .. 231
- **9.2.1. Introduction** .. 231
- **9.2.2. Robotic sample loaders** .. 231
- **9.2.3. Conclusion** .. 233

9.3. X-ray diffraction imaging of whole cells (D. Shapiro) .. 234
- **9.3.1. Introduction** .. 234
- **9.3.2. Phase retrieval from single-particle diffraction data** .. 235
- **9.3.3. High-resolution imaging of yeast** .. 235
- **9.3.4. Conclusions** .. 237

PART 10. CRYOCRYSTALLOGRAPHY .. 241

10.1. Introduction to cryocrystallography (H. Hope and S. Parkin) .. 241

- **10.1.1. Cooling of biocrystals** .. 241
- **10.1.2. Beneficial effects of low temperature** .. 242
- **10.1.3. Principles of cooling equipment** .. 244
- **10.1.4. Operational considerations** .. 245
- **10.1.5. Concluding note** .. 247

10.2. Cryocrystallography techniques and devices (D. W. Rodgers) .. 249

- **10.2.1. Introduction** .. 249
- **10.2.2. Crystal preparation** .. 249
- **10.2.3. Crystal mounting** .. 250
- **10.2.4. Flash cooling** .. 252
- **10.2.5. Transfer and storage** .. 253

10.3. Radiation damage (E. F. Garman) .. 256

- **10.3.1. Introduction** .. 256
- **10.3.2. Cryocrystallography as a mitigation strategy** .. 256
- **10.3.3. Characteristics of radiation damage at cryotemperatures** .. 257
- **10.3.4. Understanding radiation damage** .. 258
- **10.3.5. Mitigating and correcting for radiation damage** .. 259
- **10.3.6. Using radiation damage** .. 260
- **10.3.7. Open questions** .. 260

PART 11. DATA PROCESSING .. 263

11.1. Automatic indexing of oscillation images (M. G. Rossmann) .. 263

- **11.1.1. Introduction** .. 263
- **11.1.2. The crystal orientation matrix** .. 263
- **11.1.3. Fourier analysis of the reciprocal-lattice vector distribution when projected onto a chosen direction** .. 263
- **11.1.4. Exploring all possible directions to find a good set of basis vectors** .. 264
- **11.1.5. The program** .. 265

11.2. Integration of macromolecular diffraction data (A. G. W. Leslie) .. 266

- **11.2.1. Introduction** .. 266
- **11.2.2. Prerequisites for accurate integration** .. 266
- **11.2.3. Methods of integration** .. 266
- **11.2.4. The measurement box** .. 267
- **11.2.5. Integration by simple summation** .. 267
- **11.2.6. Integration by profile fitting** .. 268

11.3. Integration, scaling, space-group assignment and post refinement (W. Kabsch) .. 272

- **11.3.1. Introduction** .. 272
- **11.3.2. Modelling rotation images** .. 272
- **11.3.3. Integration** .. 275
- **11.3.4. Scaling** .. 277
- **11.3.5. Post refinement** .. 277
- **11.3.6. Space-group assignment** .. 278

11.4. *DENZO* and *SCALEPACK* (Z. Otwinowski, W. Minor, D. Borek and M. Cymborowski) .. 282

- **11.4.1. Introduction** .. 282
- **11.4.2. Diffraction from a perfect crystal lattice** .. 282
- **11.4.3. Autoindexing** .. 283
- **11.4.4. Coordinate systems** .. 284

11.4.5. Experimental assumptions 285

11.4.6. Prediction of the diffraction pattern 288

11.4.7. Integration of diffraction maxima by profile fitting 289

11.4.8. Scaling – multiplicative corrections 289

11.4.9. Global refinement or post refinement 291

11.4.10. Merging – assessment of the error model and signal magnitudes in the data 291

11.4.11. Detector diagnostics 292

11.4.12. *HKL-2000* and *HKL-3000* 292

11.4.13. Final note 294

11.5. The use of partially recorded reflections for post refinement, scaling and averaging X-ray diffraction data (C. G. van Beek, R. Bolotovsky and M. G. Rossmann) 296

11.5.1. Introduction 296

11.5.2. Generalization of the Hamilton, Rollett and Sparks equations to take into account partial reflections 296

11.5.3. Selection of reflections useful for scaling 297

11.5.4. Restraints and constraints 297

11.5.5. Generalization of the procedure for averaging reflection intensities 298

11.5.6. Estimating the quality of data scaling and averaging 298

11.5.7. Experimental results 298

11.5.8. Conclusions 301

Appendix A11.5.1. Partiality model (Rossmann, 1979; Rossmann *et al.*, 1979) 301

11.6. *XDS* (W. Kabsch) 304

11.6.1. Functional specification 304

11.6.2. *XDS* 304

11.6.3. *XSCALE* 308

11.6.4. *XDSCONV* 309

11.6.5. Parallelization of *XDS* 309

11.6.6. Availability 310

11.7. Detecting twinning by merohedry (T. O. Yeates and Y. Tsai) 311

11.7.1. Introduction 311

11.7.2. Twinning by merohedry – considerations of lattice symmetry 311

11.7.3. Considerations of length scale and effects in reciprocal space 312

11.7.4. Extent of twinning: the twin fraction 312

11.7.5. Indications of twinning 312

11.7.6. Twinning tests based on overall intensity statistics 313

11.7.7. Tests for partial twinning based on comparison of twin-related reflections 314

11.7.8. Higher forms of twinning 315

11.7.9. Other kinds of disorder 315

11.7.10. Summary 315

PART 12. ISOMORPHOUS REPLACEMENT 317

12.1. The preparation of heavy-atom derivatives of protein crystals for use in multiple isomorphous replacement and anomalous scattering (D. Carvin, S. A. Islam, M. J. E. Sternberg and T. L. Blundell) 317

12.1.1. Introduction 317

12.1.2. Heavy-atom data bank 317

12.1.3. Properties of heavy-atom compounds and their complexes 318

12.1.4. Amino acids as ligands 320

12.1.5. Protein chemistry of heavy-atom reagents 321

12.1.6. Metal-ion replacement in metalloproteins 324

12.1.7. Analogues of amino acids 324

12.1.8. Use of the heavy-atom data bank to select derivatives 325

12.2. Locating heavy-atom sites (M. T. Stubbs and R. Huber) 327

12.2.1. The origin of the phase problem 327

12.2.2. The Patterson function 328

12.2.3. The difference Fourier 329

12.2.4. Reality 329

12.2.5. Special complications 330

PART 13. MOLECULAR REPLACEMENT 333

13.1. Noncrystallographic symmetry (D. M. Blow) 333

13.1.1. Introduction 333

13.1.2. Definition of noncrystallographic symmetry 333

13.1.3. Use of the Patterson function to interpret noncrystallographic symmetry 333

13.1.4. Interpretation of generalized noncrystallographic symmetry where the molecular structure is partially known 335

13.1.5. The power of noncrystallographic symmetry in structure analysis 336

13.2. Rotation functions (J. Navaza) 340

13.2.1. Overview 340

13.2.2. Rotations in three-dimensional Euclidean space 340

13.2.3. The rotation function 341

13.2.4. The locked rotation function 343

13.2.5. Other rotation functions 344

13.2.6. Concluding remarks 344

Appendix A13.2.1. Formulae for the derivation and computation of the fast rotation function 344

13.3. Translation functions (L. Tong) 347

13.3.1. Introduction 347

13.3.2. *R*-factor and correlation-coefficient translation functions 347

13.3.3. Patterson-correlation translation function 348

13.3.4. Phased translation function 349

13.3.5. Packing check in translation functions 349

13.3.6. The unique region of a translation function (the Cheshire group) 349

13.3.7. Combined molecular replacement 349

13.3.8. The locked translation function 350

13.3.9. Miscellaneous translation functions 350

13.4. Noncrystallographic symmetry averaging of electron density for molecular-replacement phase refinement and extension (M. G. Rossmann and E. Arnold) 352

13.4.1. Introduction 352

13.4.2. Noncrystallographic symmetry (NCS) 352

13.4.3. Phase determination using NCS 354

13.4.4. The *p*- and *h*-cells 354

13.4.5. Combining crystallographic and noncrystallographic symmetry 355

13.4.6. Determining the molecular envelope 356

13.4.7. Finding the averaged density 357

13.4.8. Interpolation 358

13.4.9. Combining different crystal forms 358

13.4.10. Phase extension and refinement of the NCS parameters 359

13.4.11. Convergence 359

13.4.12. *Ab initio* phasing starts 360

13.4.13. Recent salient examples in low-symmetry cases: multidomain averaging and systematic applications of multiple-crystal-form averaging 360

13.4.14. Programs 361

13.5. Molecular replacement with *MOLREP* (A. Vagin and A. Teplyakov) 364

13.5.1. Introduction 364

13.5.2. Program operation 364

13.5.3. Preparation of the search model 364

13.5.4. Preparation of the X-ray data 364

13.5.5. Rotational search 365

13.5.6. Positional search 365

13.5.7. Multi-copy search 365

13.5.8. Fitting the model into electron density 366

13.5.9. Distribution 366

PART 14. ANOMALOUS DISPERSION 367

14.1. Heavy-atom location and phase determination with single-wavelength diffraction data (B. W. Matthews) 367

14.1.1. Introduction 367

14.1.2. The isomorphous-replacement method 367

14.1.3. The method of multiple isomorphous replacement 368

14.1.4. The method of Blow & Crick 368

14.1.5. The best Fourier 369

14.1.6. Anomalous scattering 369

14.1.7. Theory of anomalous scattering 369

14.1.8. The phase probability distribution for anomalous scattering 370

14.1.9. Anomalous scattering without isomorphous replacement 371

14.1.10. Location of heavy-atom sites 371

14.1.11. Use of anomalous-scattering data in heavy-atom location 371

14.1.12. Use of difference Fourier syntheses 371

14.1.13. Single isomorphous replacement 371

14.2. Multiwavelength anomalous diffraction (J. L. Smith and W. A. Hendrickson) 373

14.2.1. Anomalous scattering factors 373

14.2.2. A phase equation for MAD 374

14.2.3. Diffraction ratios for estimating the MAD phasing signal 375

14.2.4. Experimental considerations 375

14.2.5. Data handling 376

14.2.6. Approaches to MAD phasing 376

14.2.7. Determination of the anomalous-scatterer partial structure 377

14.2.8. General anomalous-scatterer labels for biological macromolecules 377

14.3. Automated MAD and MIR structure solution (T. C. Terwilliger and J. Berendzen) 379

14.3.1. Introduction 379

14.3.2. MAD and MIR structure solution 379

14.3.3. Decision making and structure solution 379

14.3.4. The need for rapid refinement and phasing during automated structure solution 379

14.3.5. Conversion of MAD data to a pseudo-SIRAS form 379

14.3.6. Scoring of trial heavy-atom solutions 380

14.3.7. Automated MIR and MAD structure determination 381

14.3.8. Generation of model X-ray data sets 382

14.3.9. Conclusions 382

14.3.10. Software availability 382

PART 15. DENSITY MODIFICATION AND PHASE COMBINATION 385

15.1. Phase improvement by iterative density modification (K. Y. J. Zhang, K. D. Cowtan and P. Main) 385

15.1.1. Introduction 385
15.1.2. Density-modification methods 385
15.1.3. Reciprocal-space interpretation of density modification 393
15.1.4. Phase combination 394
15.1.5. Combining constraints for phase improvement 396
15.1.6. Statistical density-modification methods 398
15.1.7. Example 398

15.2. Model phases: probabilities, bias and maps (R. J. Read) 401

15.2.1. Introduction 401
15.2.2. Model bias: importance of phase 401
15.2.3. Structure-factor probability relationships 401
15.2.4. Figure-of-merit weighting for model phases 404
15.2.5. Map coefficients to reduce model bias 404
15.2.6. Estimation of overall coordinate error 404
15.2.7. Difference-map coefficients 405
15.2.8. Refinement bias 405

15.3. *DM/DMMULTI* software for phase improvement by density modification (K. D. Cowtan, K. Y. J. Zhang and P. Main) 407

15.3.1. Introduction 407
15.3.2. Program operation 407
15.3.3. Preparation of input data 407
15.3.4. Choice of modes 408
15.3.5. Code description 410

PART 16. DIRECT METHODS 413

16.1. *Ab initio* phasing (G. M. Sheldrick, C. J. Gilmore, H. A. Hauptman, C. M. Weeks, R. Miller and I. Usón) 413

16.1.1. Introduction 413
16.1.2. Normalized structure-factor magnitudes 415
16.1.3. Starting the phasing process 416
16.1.4. Reciprocal-space phase refinement or expansion (*shaking*) 417
16.1.5. Real-space constraints (*baking*) 418
16.1.6. Fourier refinement 419
16.1.7. Resolution enhancement: the ‘free lunch’ algorithm 419
16.1.8. Utilizing Pattersons for better starts 419
16.1.9. *Shake-and-Bake*: an analysis of a dual-space method in action 420
16.1.10. Applying dual-space programs successfully 422
16.1.11. Substructure solution for native sulfurs and halide soaks 425
16.1.12. Computer programs for dual-space phasing 426
16.1.13. Conclusions and the grand challenge 429

16.2. The maximum-entropy method (G. Bricogne) 433

16.2.1. Introduction 433
16.2.2. The maximum-entropy principle in a general context 433
16.2.3. Adaptation to crystallography 435

16.3. *Ab initio* phasing of low-resolution Fourier syntheses (V. Y. Lunin, A. G. Urzhumtsev and A. Podjarny) 437

16.3.1. Introduction 437
16.3.2. General features of low-resolution images 437
16.3.3. Low-resolution phasing 437

16.3.4. Phase generation and selection 438
16.3.5. Processing of the output 439
16.3.6. Conclusions and examples 441

PART 17. MODEL BUILDING AND COMPUTER GRAPHICS 443

17.1. Macromolecular model building and validation using *Coot* (P. Emsley, B. Lohkamp and K. Cowtan) 443
17.1.1. Introduction 443
17.1.2. Model building 443
17.1.3. Validation 445
17.1.4. Scripting 446
17.1.5. Discussion 446

17.2. Molecular graphics and animation (A. J. Olson) 448
17.2.1. Introduction 448
17.2.2. Background – the evolution of molecular graphics hardware and software 448
17.2.3. Representation and visualization of molecular data and models 449
17.2.4. Presentation graphics 454
17.2.5. Looking ahead 457

PART 18. REFINEMENT 459

18.1. Introduction to refinement (L. F. Ten Eyck and K. D. Watenpaugh) 459
18.1.1. Overview 459
18.1.2. Background 459
18.1.3. Objectives 459
18.1.4. Least squares and maximum likelihood 460
18.1.5. Optimization 460
18.1.6. Data 460
18.1.7. Models 461
18.1.8. Optimization methods 462
18.1.9. Evaluation of the model 464
18.1.10. Conclusion 464

18.2. Enhanced macromolecular refinement by simulated annealing (A. T. Brunger, P. D. Adams and L. M. Rice) 466
18.2.1. Introduction 466
18.2.2. Cross validation 466
18.2.3. The target function 467
18.2.4. Searching conformational space 468
18.2.5. Examples 470
18.2.6. Multi-start refinement and structure-factor averaging 471
18.2.7. Ensemble models 471
18.2.8. Conclusions 472

18.3. Structure quality and target parameters (R. A. Engh and R. Huber) 474
18.3.1. Purpose of restraints 474
18.3.2. Formulation of refinement restraints 475
18.3.3. Strategy of application during building/refinement 483
18.3.4. Future perspectives 483

18.4. Refinement at atomic resolution (Z. Dauter, G. N. Murshudov and K. S. Wilson) 485
18.4.1. The atomic model and a definition of atomic resolution 485
18.4.2. Data 487
18.4.3. Computational algorithms and strategies 488
18.4.4. Computational options and tactics 489

18.4.5. Features in the refined model 491

18.4.6. Quality assessment of the model 495

18.4.7. Relation to biological chemistry 495

18.4.8. Practical strategies 496

18.5. Coordinate uncertainty (D. W. J. Cruickshank) 499

18.5.1. Introduction 499

18.5.2. The least-squares method 500

18.5.3. Restrained refinement 501

18.5.4. Two examples of full-matrix inversion 502

18.5.5. Approximate methods 505

18.5.6. The diffraction-component precision index 506

18.5.7. Examples of the diffraction-component precision index 507

18.5.8. Luzzati plots 509

18.6. *CNS*, a program system for structure-determination and refinement (A. T. Brunger, P. D. Adams, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, N. S. Pannu, R. J. Read, L. M. Rice and T. Simonson) 512

18.6.1. Introduction 512

18.6.2. The CNS language 512

18.6.3. Symbols and parameters 513

18.6.4. Statistical functions 513

18.6.5. Symbolic target function 514

18.6.6. Modules and procedures 515

18.6.7. Task files 516

18.6.8. HTML interface 516

18.6.9. Example: combined maximum-likelihood and simulated-annealing refinement 517

18.6.10. Conclusions 517

18.7. The *TNT* refinement package (D. E. Tronrud and L. F. Ten Eyck) 520

18.7.1. Scope and function of the package 520

18.7.2. Historical context 520

18.7.3. Design principles 520

18.7.4. Current structure of the package 522

18.7.5. Innovations first introduced in *TNT* 522

18.7.6. *TNT* as a research tool 523

18.7.7. Current status of *TNT* 523

18.8. *ARP/wARP* – automated model building and refinement (V. S. Lamzin, A. Perrakis and K. S. Wilson) 525

18.8.1. Refinement and model building are two parts of modelling a structure 525

18.8.2. Free-atom and hybrid models 525

18.8.3. *ARP/wARP* applications 526

18.8.4. Iterations 528

18.8.5. Applicability and requirements 528

18.9. Macromolecular applications of *SHELX* (G. M. Sheldrick) 529

18.9.1. Introduction 529

18.9.2. Experimental phasing with *SHELXC/D/E* 529

18.9.3. Macromolecular refinement using *SHELXL* 531

18.9.4. *SHELXPRO* – protein interface to *SHELX* 532

18.9.5. Distribution and support of *SHELX* 532

18.10. *PrimeX* and the Schrödinger computational chemistry suite of programs (J. A. Bell, Y. Cao, J. R. Gunn, T. Day, E. Gallicchio, Z. Zhou, R. Levy and R. Farid) 534

18.10.1. Introduction 534

18.10.2. Computational environment 534

18.10.3. Achieving the mission of *PrimeX* 535

18.10.4. *PrimeX* implementation and theory 536

18.10.5. Conclusion 538

18.11. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution (P. D. ADAMS, P. V. AFONINE, G. BUNKÓCZI, V. B. CHEN, I. W. DAVIS, N. ECHOLS, J. J. HEADD, L.-W. HUNG, G. J. KAPRAL, R. W. GROSSE-KUNSTLEVE, A. J. MCCOY, N. W. MORIARTY, R. OEFFNER, R. J. READ, D. C. RICHARDSON, J. S. RICHARDSON, T. C. TERWILLIGER AND P. H. ZWART) 539

18.11.1. Foundations 539

18.11.2. Analysis of experimental data 540

18.11.3. Substructure determination, phasing and molecular replacement 540

18.11.4. Model building, ligand fitting and nucleic acids 541

18.11.5. Model, and model-to-data, validation 542

18.11.6. Structure refinement 543

18.11.7. Integrated structure determination 544

18.11.8. Conclusions 545

18.12. Structure determination in the presence of twinning by merohedry (T. O. YEATES AND M. R. SAWAYA) 548

18.12.1. Introduction 548

18.12.2. Detwinning based on observed intensities 548

18.12.3. Molecular replacement with twinning 548

18.12.4. Multiple isomorphous replacement and anomalous phasing with twinning 548

18.12.5. Atomic refinement with twinning 549

18.12.6. Detwinning on the basis of model $F_{\rm calc}$ values 550

18.12.7. Summary 550

PART 19. OTHER EXPERIMENTAL TECHNIQUES 553

19.1. Neutron crystallography: methods and information content (A. A. KOSSIAKOFF) 553

19.1.1. Introduction 553

19.1.2. Diffraction geometries 553

19.1.3. Neutron density maps – information content 553

19.1.4. Phasing models and evaluation of correctness 554

19.1.5. Evaluation of correctness 554

19.1.6. Refinement 555

19.1.7. $D_2O - H_2O$ solvent difference maps 555

19.1.8. Applications of $D_2O - H_2O$ solvent difference maps 555

19.2. Electron diffraction of protein crystals (W. CHIU) 557

19.2.1. Electron scattering 557

19.2.2. The electron microscope 557

19.2.3. Data collection 557

19.2.4. Data processing 558

19.2.5. Future development 561

19.3. Small-angle X-ray scattering (H. TSURUTA AND J. E. JOHNSON) 563

19.3.1. Introduction 563

19.3.2. Small-angle single-crystal X-ray diffraction studies 563

19.3.3. Solution X-ray scattering studies 564

19.4. Small-angle neutron scattering (D. M. ENGELMAN AND P. B. MOORE) 575

19.4.1. Introduction 575

19.4.2. Fundamental relationships 575

19.4.3. Contrast variation 577

19.4.4. Distance measurements 579

19.4.5. Practical considerations 580

19.4.6. Examples 580

19.5. Fibre diffraction (R. Chandrasekaran and G. Stubbs) 583

19.5.1. Introduction 583

19.5.2. Types of fibres 583

19.5.3. Diffraction by helical molecules 584

19.5.4. Fibre preparation 585

19.5.5. Data collection 585

19.5.6. Data processing 585

19.5.7. Determination of structures 586

19.5.8. Structures determined by X-ray fibre diffraction 588

19.6. Electron cryomicroscopy of biological macromolecules (T. S. Baker and R. Henderson) 593

19.6.1. Abbreviations used 593

19.6.2. Introduction: macromolecular structure determination using electron microscopy 593

19.6.3. Physics of electron scattering and radiation damage 593

19.6.4. Three-dimensional electron cryomicroscopy of macromolecules 595

19.6.5. Image processing and 3D reconstruction 601

19.6.6. Visualization, modelling and interpretation of results 604

19.6.7. Trends 605

19.7. Nuclear magnetic resonance (NMR) spectroscopy (K. Wüthrich) 615

19.7.1. Complementary roles of NMR in solution and X-ray crystallography in structural biology 615

19.7.2. A standard protocol for NMR structure determination of proteins and nucleic acids 615

19.7.3. Combined use of single-crystal X-ray diffraction and solution NMR for structure determination 617

19.7.4. NMR studies of solvation in solution 617

19.7.5. NMR studies of rate processes and conformational equilibria in three-dimensional macromolecular structures 618

19.8. Use of *SPIDER* and *SPIRE* in image reconstruction (A. Leith, W. Baxter and J. Frank) 620

19.8.1. Introduction 620

19.8.2. Basic philosophy of single-particle reconstruction 620

19.8.3. Implementation of single-particle reconstruction in *SPIDER* 621

19.8.4. Implementation of single-particle reconstruction in *SPIRE* 623

19.8.5. Conclusion 623

19.9. Four-dimensional cryo-electron microscopy at quasi-atomic resolution: *IMAGIC 4D* (M. van Heel, R. Portugal, A. Rohou, C. Linnemayr, C. Bebeacua, R. Schmidt, T. Grant and M. Schatz) 624

19.9.1. Introduction 624

19.9.2. The *IMAGIC* software system 624

19.9.3. *IMAGIC* '4D' processing/data format 624

19.9.4. Software parallelization 625

19.9.5. Full 2D (parallel) astigmatic contrast transfer function correction 625

19.9.6. Parallel automatic particle picking 625

19.9.7. Parallel multi-reference alignments and reference bias 626

19.9.8. MSA and its parallelization 626

19.9.9. Handling multiple 3D reconstructions in parallel 626

19.9.10. Angular reconstitution 4D refinements 627

19.9.11. Discussion 627

19.10. Single-particle reconstruction with *EMAN* (S. Ludtke) 629

19.10.1. Introduction 629

19.10.2. Overview of *EMAN* 629

19.10.3. Single-particle reconstruction 631

19.10.4. Evaluating the reconstruction 631

PART 20. ENERGY CALCULATIONS AND MOLECULAR DYNAMICS 633

20.1. Molecular-dynamics simulation of protein crystals: convergence of molecular properties of ubiquitin (U. Stocker and W. F. van Gunsteren) 633

20.1.1. Introduction 633

20.1.2. Methods 633

20.1.3. Results 634

20.1.4. Conclusions 640

20.2. Molecular-dynamics simulations of biological macromolecules (C. B. Post and V. M. Dadarlat) 642

20.2.1. Introduction 642

20.2.2. The simulation method 642

20.2.3. Potential-energy function 642

20.2.4. Empirical parameterization of the force field 644

20.2.5. Modifications in the force field for structure determination 644

20.2.6. Internal dynamics and average structures 644

20.2.7. Assessment of the simulation procedure 645

20.2.8. Effect of crystallographic atomic resolution on structural stability during molecular dynamics 645

PART 21. STRUCTURE VALIDATION 649

21.1. Validation of protein crystal structures (G. J. Kleywegt) 649

21.1.1. Introduction 649

21.1.2. Types of error 649

21.1.3. Detecting outliers 650

21.1.4. Fixing errors 651

21.1.5. Preventing errors 652

21.1.6. Final model 652

21.1.7. A compendium of quality criteria 652

21.1.8. Future 658

21.2. Assessing the quality of macromolecular structures (S. J. Wodak, A. A. Vagin, J. Richelle, U. Das, J. Pontius and H. M. Berman) 662

21.2.1. Introduction 662

21.2.2. Validating the geometric and stereochemical parameters of the model 662

21.2.3. Validation of a model *versus* experimental data 665

21.2.4. Atomic resolution structures 673

21.2.5. Concluding remarks 673

21.3. Detection of errors in protein models (O. Dym, D. Eisenberg and T. O. Yeates) 677

21.3.1. Motivation and introduction 677

21.3.2. Separating evaluation from refinement 677

21.3.3. Algorithms for the detection of errors in protein models and the types of errors they detect 677

21.3.4. Selection of database 679

21.3.5. Examples: detection of errors in structures 679

21.3.6. Summary 682

21.3.7. Availability of software 682

21.4. *PROCHECK*: validation of protein-structure coordinates (R. A. Laskowski, M. W. MacArthur and J. M. Thornton) 684

21.4.1. Introduction 684

21.4.2. The program 684

21.4.3. The parameters 684

21.4.4. Which parameters are best? 685

21.4.5. Input 686

21.4.6. Output produced 686

21.4.7. Other validation tools .. 686

21.5. *KiNG* and kinemages (V. B. Chen, J. S. Richardson and D. C. Richardson) .. 688

21.5.1. Introduction to aims and concepts .. 688

21.5.2. Uses of *KiNG* and kinemages .. 689

21.5.3. Making kinemages .. 692

21.5.4. Software notes .. 693

21.6. *MolProbity*: all-atom structure validation for macromolecular crystallography (V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson) .. 694

21.6.1. Summary of *MolProbity* flow and user interactions .. 694

21.6.2. Validation analyses .. 694

21.6.3. Correction of outliers .. 698

21.6.4. Other *MolProbity* utility functions .. 699

21.6.5. Discussion .. 700

21.6.6. *MolProbity* availability .. 701

PART 22. MOLECULAR GEOMETRY AND FEATURES .. 703

22.1. Protein geometry: volumes, areas and distances (M. Gerstein and F. M. Richards) .. 703

22.1.1. Introduction .. 703

22.1.2. Definitions of protein volume .. 703

22.1.3. Definitions of protein surface .. 706

22.1.4. Definitions of atomic radii .. 708

22.1.5. Application of geometry calculations: the measurement of packing .. 709

22.2. Molecular surfaces: calculations, uses and representations (M. S. Chapman and M. L. Connolly) .. 713

22.2.1. Introduction .. 713

22.2.2. Calculation of surface area and energies of interaction .. 714

22.2.3. Estimation of binding energies .. 715

22.2.4. Graphical representations of shape and properties .. 717

22.2.5. Conclusion .. 719

22.3. Hydrogen bonding in biological macromolecules (E. N. Baker) .. 721

22.3.1. Introduction .. 721

22.3.2. Nature of the hydrogen bond .. 721

22.3.3. Hydrogen-bonding groups .. 721

22.3.4. Identification of hydrogen bonds: geometrical considerations .. 722

22.3.5. Hydrogen bonding in proteins .. 723

22.3.6. Hydrogen bonding in nucleic acids .. 726

22.3.7. Non-conventional hydrogen bonds .. 727

22.4. Electrostatic interactions in proteins (K. A. Sharp) .. 730

22.4.1. Introduction .. 730

22.4.2. Theory .. 730

22.4.3. Applications .. 732

22.5. The relevance of the Cambridge Structural Database in protein crystallography (F. H. Allen, J. C. Cole and M. L. Verdonk) 736

22.5.1. Introduction .. 736

22.5.2. The CSD and the PDB: data acquisition and data quality .. 736

22.5.3. Structural knowledge from the CSD .. 737

22.5.4. Intramolecular geometry .. 738

22.5.5. Intermolecular data .. 740

22.5.6. Conclusion .. 745

PART 23. STRUCTURAL ANALYSIS AND CLASSIFICATION 749

23.1. Protein-fold classification (C. Orengo and J. Thornton) 749

23.2. Locating domains in three-dimensional structures (L. Holm and C. Sander) 752

23.2.1. Introduction 752

23.2.2. Compactness 752

23.2.3. Recurrence 752

23.2.4. Conclusion 753

23.3. Protein–ligand interactions (A. E. Hodel and F. A. Quiocho) 755

23.3.1. Introduction 755

23.3.2. Protein–carbohydrate interactions 755

23.3.3. Metals 756

23.3.4. Protein–nucleic acid interactions 757

23.3.5. Phosphate and sulfate 761

23.4. Nucleic acids (R. E. Dickerson) 766

23.4.1. Introduction 766

23.4.2. Helix parameters 766

23.4.3. Comparison of A, B and Z helices 773

23.4.4. Sequence–structure relationships in B-DNA 779

23.4.5. Summary 784

Appendix A23.4.1. X-ray analyses of A, B and Z helices 787

23.5. Solvent structure (C. Mattos and D. Ringe) 800

23.5.1. Introduction 800

23.5.2. Determination of water molecules 801

23.5.3. Structural features of protein–water interactions derived from database analysis 802

23.5.4. Water structure in groups of well studied proteins 808

23.5.5. The classic models: small proteins with high-resolution crystal structures 814

23.5.6. Water molecules as mediators of complex formation 815

23.5.7. Conclusions and future perspectives 817

23.6. Halogen interactions in biomolecular crystal structures (M. J. Vallejos, P. Auffinger and P. S. Ho) 821

23.6.1. Introduction 821

23.6.2. Classical treatment of halogen interactions 821

23.6.3. Electrostatic molecular halogen interactions 822

23.6.4. Concluding remarks 825

PART 24. CRYSTALLOGRAPHIC DATABASES 827

24.1. The Worldwide Protein Data Bank (H. M. Berman, K. Henrick, G. Kleywegt, H. Nakamura and J. Markley) 827

24.1.1. Introduction 827

24.1.2. Data acquisition and processing 827

24.1.3. Data access 829

24.1.4. Future 831

24.2. The Nucleic Acid Database (B. Schneider, J. de la Cruz, S. Dutta, Z. Feng, L. Chen, J. Westbrook, H. Yang, J. Young, C. Zardecki and H. M. Berman) 833

24.2.1. Introduction 833

24.2.2. Data processing and validation 833

24.2.3. The database 834

24.2.4. Distribution of information 835

24.3. The Biological Macromolecule Crystallization Database (D. T. Gallagher and M. Tung) 838

24.3.1. Introduction 838

24.3.2. History of the BMCD 838

24.3.3. BMCD data 838

24.3.4. Web interface 839

24.3.5. Reproducing published crystallization procedures 839

24.3.6. Crystallization screens 840

24.3.7. A general crystallization procedure 840

24.3.8. The future of the BMCD 841

PART 25. A HISTORICAL PERSPECTIVE 845

25.1. How the structure of lysozyme was actually determined (C. C. F. Blake, R. H. Fenn, L. N. Johnson, D. F. Koenig, G. A. Mair, A. C. T. North, J. W. H. Oldham, D. C. Phillips, R. J. Poljak, V. R. Sarma and C. A. Vernon) 845

25.1.1. Introduction 845

25.1.2. Structure analysis at 6 Å resolution 845

25.1.3. Analysis of the structure at 2 Å resolution 854

25.1.4. Structural studies on the biological function of lysozyme 866

Subject index 874

Preface

E. ARNOLD AND M. G. ROSSMANN

International Tables for Crystallography, Volume F, *Crystallography of Biological Macromolecules*, was commissioned by the International Union of Crystallography (IUCr) in recognition of the extraordinary contributions that knowledge of macromolecular structure has made, and will make, to the analysis of biological systems, from enzyme catalysis to the workings of a whole cell. The volume covers all stages of a crystallographic analysis from the preparation of samples using the techniques of molecular biology and biochemistry, to crystallization, diffraction-data collection, phase determination, structure validation and structure analysis. Although the book is written for experienced scientists, it is recognized that the modern structural biologist is more likely to be a biologist interested in structure than a classical crystallographer interested in biology. Thus, there are chapters on the fundamentals, history and current perspectives of macromolecular crystallography, as well as on the availability of useful programs and databases including the Protein Data Bank. Each chapter has been written by an internationally recognized expert.

Macromolecular crystallography is undergoing a revolution. Just as crystallography became central to the study of chemistry, macromolecular crystallography has become a core science in biology. Macromolecular crystallography has shaped our view of biological molecular structure, and is providing a broader understanding of biological ultrastructure and the molecular interactions in living systems. As reflected by the exponential increase in entries in the Protein Data Bank over the past decade, there has been an explosion in the number of macromolecular structures determined, the majority by X-ray crystallography. Knowledge of the sequences of entire genomes, from bacteria to human, has sparked a structural genomics effort that aims to determine 10 000 new macromolecular structures in the next decade. Crystallography is expected to yield the largest share of this new crop of structures. The field of macromolecular crystallography is still evolving rapidly, and capturing its essence in a single volume is a challenge. Therefore, the volume emphasizes durable knowledge, but also contains articles on somewhat more volatile topics.

This project had its inception when Ted Baker (at that time President of the IUCr) approached one of us (MGR) about writing a book on macromolecular crystallography for the IUCr. Not only were there already some excellent books that covered most aspects of the subject, but the breadth of the subject was now so vast that no single person could possibly be an expert in all relevant topics. After further exchanges of e-mails, MGR realized that the officers of the IUCr were tacitly assuming that he would be willing to carry out the advice he had given so freely. He then asked his former post-doc and coauthor of an earlier article on molecular replacement in Volume B of *International Tables*, Eddy Arnold, to help him get out of a tight corner. After some serious deliberations of his own, Eddy agreed to be co-editor.

Together we fleshed out an outline that was broader than MGR's original plan, which had focused largely on crystallographic theory and technique. We felt that it would be valuable to briefly cover related techniques beyond X-ray diffraction, as well as to give an overview of the current field of structural biology. Although basic crystallography is also presented in the other volumes of *International Tables*, chapters describing fundamental crystallographic principles and practices have been included in an attempt to make the volume as coherent and self-contained as possible. We established an advisory board, developed a list of required chapters and obtained promises of participation from potential authors. In a departure from the style of previous volumes of *International Tables*, which have fewer articles and authors, we sought contributions for nearly 100 articles from an even larger number of contributing authors. The members of the advisory board reviewed the proposed outline of chapters and authors. We were pleasantly surprised when so many experts generously agreed to write articles for this volume, and delighted that the vast majority fulfilled their promises.

Significant events punctuating the process were the 1996 and 1999 IUCr congresses. At the 1996 IUCr Congress in Seattle, we convened a meeting with many of the authors. There we described the overall project design and received valuable suggestions. At that time, we hoped that the volume could be completed by 1999. At the 1999 IUCr Congress in Glasgow, we reviewed the detailed contents of the volume at an open meeting on the volumes of *International Tables* under development. By that time, we had received most of the articles and typesetting began in late 1999. The complexities of handling a large number of articles from so many authors led to delays at a number of stages. Ultimately, the completion date became mid-2001.

We are especially grateful to the staff at the IUCr and at our own institutions for their dedicated help in bringing this project to fruition. At the IUCr, we thank Nicola Ashcroft for an outstanding job on overall production of the volume, and for her patient correspondence and attention to detail. We also thank Peter Strickland, Sue King, Theo Hahn, Uri Shmueli, Mike Dacombe and Ted Baker for their help in coordinating the project. At Purdue University, we thank Cheryl Towell and Sharon Wilder for constant assistance, and Fay Chen for editorial suggestions. At the Center for Advanced Biotechnology and Medicine and Rutgers University, we thank Susan Mazzocchi and Barbara Shaver for their help in handling correspondence and galley proofs from the authors.

We are also especially indebted to the authors for their generous contributions and for documenting relevant expertise. We also thank the advisors and the members of the advisory board for their help. We are saddened to note that Paul Sigler, a member of the advisory board, passed away during the project. Paul was a towering figure who, with his medical background, recognized the role structure plays in providing insights into fundamental chemical and biological processes.

Preface to the second edition

E. Arnold, D. M. Himmel and M. G. Rossmann

Ten years after the appearance of the first edition of *International Tables for Crystallography*, Volume F, *Crystallography of Biological Macromolecules*, we are pleased to present the second edition. Preparation of the second edition has reflected the continuing evolution of macromolecular crystallography. Eddy Arnold and Michael Rossmann were glad to have Daniel Himmel, an accomplished crystallographer with diverse knowledge, join this effort. The three of us brainstormed about how to update this volume most effectively, and decided to seek new articles in key areas of rapid growth and ask authors from the first edition if they wished to revise their chapters. We were delighted when so many internationally recognized experts again made generous contributions that brought the volume up to date.

In the past decade the field of macromolecular crystallography has spawned an ever-increasing array of spectacular biological structures. The amazing conquests have included the translational machinery, the ribosome, in various functional states; the transcription machinery, multisubunit cellular RNA polymerases; a multitude of membrane proteins; and many more viruses. The marriage of crystallography and cryo-electron microscopy, including electron tomography, enables a multiscale view of biological structure including macromolecular inventories of cells in three dimensions. Multiprotein complexes are more routinely studied, and small-angle X-ray scattering has grown in prominence as a complementary technique to single-crystal X-ray diffraction. Structural genomics projects have been knocking out representative structures from a wide variety of proteomes from all levels of life. Structure-based drug-design prospects and the general problem of understanding protein–ligand interactions have benefitted from the development of techniques such as fragment screening, where thousands of small-molecule fragments are systematically soaked into crystals to interrogate potential binding sites – a central tool in the emerging field of chemical genetics. Software packages continue to evolve to enable rapid and reliable structure solution, visualization and analysis. X-ray lasers are now a reality and perhaps in the future many complex structures will be solved and imaged using data sets measured from tiny microcrystals that diffract their heart out before being destroyed by the pulsed high-brilliance beams. When thinking about the next ten years, we can expect more of the unexpected.

We wish to thank Nicola Ashcroft for handling the majority of correspondence with authors, and for maximizing the efficiency and quality of the overall process, thereby keeping the production schedule under control. We also thank our many colleagues and friends whose advice was valuable in designing the plan for the second edition.

International Tables for Crystallography (2012). Vol. F, Chapter 1.1, pp. 1–4.

PART 1. INTRODUCTION

Chapter 1.1. Overview

E. Arnold, D. M. Himmel and M. G. Rossmann

Volume F is devoted to the crystallography of large biological molecules, complementing the existing volumes of *International Tables for Crystallography*. A background history of the subject is followed by a concise introduction to the basic theory of X-ray diffraction and other requirements for the practice of crystallography. Basic crystallographic theory is presented in greater depth in other volumes of *International Tables*. The information in the latter portions of Volume F is specifically related to macromolecular structure.

Chapter 1.2 presents a brief history of the field of macromolecular crystallography. This is followed by an article (Chapter 1.3) that describes many of the connections of crystallography with medicine and gives a look into the future possibilities of structure-based design of drugs, vaccines and other agents. Chapter 1.4 provides some personal perspectives on what the future may hold for crystallography and the other physical sciences, as well as how budding new technologies and breakthroughs in these fields may impact both the sciences and the world at large.

Chapter 2.1 introduces diffraction theory and fundamentals of crystallography, including concepts of real and reciprocal space, unit-cell geometry and symmetry. It is shown how scattering from electron density and atoms leads to the formulation of structure factors. The phase problem is introduced, as well as the basic theory behind some of the more common methods for its solution. As the field of macromolecular crystallography has grown and matured, a variety of metrics have evolved for monitoring the quality of data collection, data processing and structure determination. In some cases, disagreement has arisen over the best way to calculate some of these metrics or which ones are most useful for monitoring the progress of the crystallographic experiment. Chapter 2.2 provides a comprehensive compilation of the most commonly used indicators of quality, along with a definition for each. The application of some of these indicators to validate the refined crystal structure is addressed further in Part 21 (below).

Molecular biology has had a major impact in terms of accelerating progress in structural biology and remains a rapidly developing area. Chapter 3.1 is a primer on modern molecular-biology techniques for producing materials for crystallographic studies. Since large amounts of highly purified materials are required, emphasis is placed on approaches for efficiently and economically yielding samples of biological macromolecules suitable for crystallization. Chapter 3.2 extends this overview to methods for the expression and purification of membrane proteins for structural studies. These introductions to protein expression are complemented by Chapter 4.3, which describes the rapidly developing area of molecular-engineering approaches for obtaining high-quality crystals of biological macromolecules.

The basic theory and practice of macromolecular crystallization are described in Chapters 4.1 and 4.2. This, too, is a rapidly evolving area, with continual advances in theory and practice. It is remarkable to consider the macromolecules that have been crystallized. We expect macromolecular engineering, covered in Chapter 4.3, to play a central role in coaxing more macromolecules to form crystals suitable for structure determination in the future. These areas will be complemented by high-throughput crystallographic approaches, covered in Chapter 4.4. These include the expression of proteins in various organisms for their structure determination. The material in Part 4 is complemented by Part 5, which summarizes traditional properties of and methods for handling macromolecular crystals, as well as how to measure crystal density.

Part 6 gives a brief introduction to the theory and practice of generating X-rays and neutrons for diffraction experiments. Chapter 6.1 describes the theory of X-ray production from both conventional and synchrotron X-ray sources, as well as methods for defining the energy spectrum and geometry of X-ray beams. Numerous articles in the other volumes of *International Tables* go into greater depths. Chapter 6.2 describes the generation of neutron beams.

Part 7 describes common methods for detecting X-rays, with a focus on detection devices that are currently most frequently used, including storage phosphor image plate and charge-coupled device (CCD) detectors. This has been another rapidly developing area, particularly in the past two decades. A further article describing X-ray detector theory and practice is *International Tables* Volume C, Chapter 7.1.

Synchrotron-radiation sources have played a prominent role in advancing the frontiers of macromolecular structure determination in terms of size, quality and throughput. The extremely high intensity, tunable wavelength characteristics and pulsed time structure of synchrotron beams have enabled many novel experiments. Some of the unique characteristics of synchrotron radiation are being harnessed to help solve the phase problem using anomalous-scattering measurements, *e.g.* in multiwavelength anomalous diffraction (MAD) experiments (see Chapter 14.2). The quality of synchrotron-radiation facilities for macromolecular studies has been increasing rapidly, partly in response to the perceived value of the structures being determined. Many synchrotron beamlines have been designed to meet the needs of macromolecular experiments. Chapter 8.1 surveys many of the roles that synchrotron radiation plays in modern macromolecular structure determination. Chapter 8.2 summarizes applications of the age-old Laue crystallography technique, which has seen a revival in the study of macromolecular crystal structures using portions of the white spectrum of synchrotron X-radiation. Chapter 4.2 of *International Tables* Volume C is also a useful reference for understanding synchrotron radiation.

Chapter 9.1 summarizes many aspects of data collection from single crystals using monochromatic X-ray beams. Common camera-geometry and coordinate-system-definition schemes are given. Because most macromolecular data collection is carried out using the oscillation (or rotation) method, strategies related to this technique are emphasized. A variety of articles in Volume C of *International Tables* serve as additional references. With the

Copyright © 2012 International Union of Crystallography

advent of modern synchrotron-radiation facilities that provide high-intensity beams and CCD detectors with rapid data readout, data collection has often become so rapid that the rate-limiting step in collecting a data set could very well be the time it takes the experimenter to mount the crystal manually, lock the hutch and begin to take X-ray exposures. The introduction of robotic crystal loading, covered in Chapter 9.2, makes it possible to mount and screen many more crystals, collecting and saving data from them, in the same amount of time, often with far less risk to the crystals. Chapter 9.3 summarizes another technology, coherent X-ray diffraction microscopy (CXDM), which is made possible by the brightness of third-generation synchrotron X-ray sources. Using this method, diffractive methods can be applied to image certain non-crystalline single particles to as high as 11 nm resolution. Single-particle phase retrieval theory is briefly presented, followed by the application of CXDM to the imaging of a yeast cell, a discussion of the ramifications of radiation damage and, finally, the use of stereoscopic viewing to access three-dimensional information.

The use of cryogenic cooling of macromolecular crystals for data collection ('cryocrystallography') has become the most frequently used method of crystal handling for data collection. Part 10 summarizes the theory and practice of cryocrystallography. Among its advantages are enhanced crystal lifetime and improved resolution. Most current experiments in cryocrystallography use liquid-nitrogen-cooled gas streams, though some attempts have been made to use liquid-helium-cooled gas streams. Although cryogenic temperatures substantially protect macromolecular crystals against radiation damage from the generation of free radicals, the problem of radiation damage has not been entirely eliminated. Methods continue to be explored to reduce the rate of radiation damage, such as the use of free-radical scavengers in cryoprotective solutions, or to correct reflections data sets for X-ray radiation dose. Methods have also been developed for utilizing radiation damage for phasing. Just a decade ago, it was still widely believed that many macromolecular crystals could not be studied successfully using cryocrystallography, or that the practice would be troublesome or would lead to inferior results. Now, crystallographers routinely screen for suitable cryoprotective conditions for data collection even in initial experiments, and often crystal diffraction quality is no longer assessed except using cryogenic cooling. However, some crystals have resisted attempts to cool them successfully to cryogenic temperatures. Thus, data collection using ambient conditions, or moderate cooling (from approximately −40 °C to a few degrees below ambient temperature), is not likely to become obsolete in the near future.

Part 11 describes the processing of X-ray diffraction data from macromolecular crystals. Special associated problems concern dealing with large numbers of observations, large unit cells (hence crowded reciprocal lattices) and diverse factors related to crystal imperfection (large and often anisotropic mosaicity, variability of unit-cell dimensions *etc.*). Various camera geometries have been used in macromolecular crystallography, including precession, Weissenberg, three- and four-circle diffractometry, and oscillation or rotation. The majority of diffraction data sets are collected now *via* the oscillation method and using a variety of detectors. Among the topics covered in Part 11 are autoindexing, intensity integration, space-group assignment, scaling, post refinement, and detection of merohedral twinning. Commonly used software packages for the indexing, integration, scaling and post refinement of crystallographic data are presented. The discussion of intensity integration in Chapter 11.2 includes a description of some of the strategies used by *MOSFLM* and *DENZO*. *DENZO* and *SCALEPACK*, which form the backbone of the *HKL2000* package, are described in detail in Chapter 11.4, and *XDS* is described in Chapter 11.6.

Part 12 describes the theory and practice of the isomorphous replacement method, and begins the portion of Volume F that addresses how the phase problem in macromolecular crystallography can be solved. The isomorphous replacement method was the first technique used for solving macromolecular crystal structures, and will continue to play a central role for the foreseeable future. Chapter 12.1 describes the basic practice of isomorphous replacement, including the selection of heavy-metal reagents as candidate derivatives and crystal-derivatization procedures. Chapter 12.2 surveys some of the techniques used in isomorphous replacement calculations, including the location of heavy-atom sites and use of that information in phasing. Readers are also referred to Chapter 2.4 of *International Tables* Volume B for additional information about the isomorphous replacement method.

Part 13 describes the molecular replacement method and many of its uses in solving macromolecular crystal structures. This part covers general definitions of noncrystallographic symmetry, the use of rotation and translation functions, and phase improvement and extension *via* noncrystallographic symmetry. The molecular replacement method is commonly used to solve macromolecular crystal structures where redundant information is present either in a given crystal lattice or among different crystals. In some cases, phase information is obtained by averaging noncrystallographically redundant electron density either within a single crystal lattice or among multiple crystal lattices. In other cases, atomic models from known structures can be used to help phase unknown crystal structures containing related structures. With the number and type of known protein structures rising exponentially in recent years (see Chapter 24.1), molecular replacement has become the most common way to solve phases for an unknown protein structure. Molecular-replacement phasing is often used in conjunction with other phasing methods, including isomorphous replacement and density-modification methods. *International Tables* Volume B, Chapter 2.3 is also a useful reference for molecular-replacement techniques.

Anomalous-dispersion measurements have played an increasingly important role in solving the phase problem for macromolecular crystals. Anomalous dispersion has long been recognized as a source of experimental phase information; for more than three decades, macromolecular crystallographers have been exploiting anomalous-dispersion measurements from crystals containing heavy metals, using even conventional X-ray sources. In the past two decades, synchrotron sources have permitted optimized anomalous-scattering experiments, where the X-ray energy is selected to be near an absorption edge of a scattering element. Chapter 14.1 summarizes applications of anomalous scattering using single wavelengths for macromolecular crystal-structure determination. The multiwavelength anomalous diffraction (MAD) technique, in particular, is used to solve the phase problem for a broad array of macromolecular crystal structures. In the MAD experiment, intensities measured from a crystal at a number of wavelengths permit direct solution of the phase problem, frequently yielding easily interpretable electron-density maps. The theory and practice of the MAD technique are described in Chapters 14.2 and 14.3.

Density modification, discussed in Part 15, encompasses an array of techniques used to aid solution of the phase problem *via* electron-density-map modifications. Recognition of usual density-distribution patterns in macromolecular crystal structures permits the application of such techniques as solvent flattening (disordered solvent regions have lower density), histogram matching (normal distributions of density are expected) and skeletonization (owing to the long-chain nature of macromolecules such as proteins). Electron-density averaging, discussed in Chapter 13.4, can be thought of as a density-modification technique as well. Chapter 15.1 surveys the general problem and practice of density modification, including a discussion of solvent flattening, histogram matching, skeletonization and phase-combination methodology. Chapter 15.2 discusses weighting of Fourier terms for calculation of electron-density maps in a more general sense, especially with respect to the problem of minimizing model bias in phase improvement. Electron-density modification techniques can often be implemented efficiently in reciprocal space, too. Chapter 15.3 summarizes the implementation of phase modification in the *DM/DMMULTI* software.

Part 16 describes the use of direct methods in macromolecular crystallography. Some 30 years ago, direct methods revolutionized the practice of small-molecule crystallography by facilitating structure solution directly from intensity measurements. As a result, phase determination of most small-molecule crystal structures has become quite routine. In the meantime, many attempts have been made to apply direct methods to solving macromolecular crystal structures. Prospects in this area are improving, but success has been obtained in only a limited number of cases, often with extremely high resolution data measured from small proteins. Chapter 16.1 surveys progress in the application of direct methods to solve macromolecular crystal structures, and Chapter 16.2 summarizes the application of the maximum-entropy method to crystallography. Chapter 16.3 explores techniques now being developed to extend *ab initio* phasing to low-resolution data.

The use of computer graphics for building models of macromolecular structures has facilitated the efficiency of macromolecular structure solution and refinement immensely (Part 17). Until just a little more than 20 years ago, all models of macromolecular structures were built as physical models, with parts of appropriate dimensions scaled up to our size! Computer-graphics representations of structures have made macromolecular structure models more precise, especially when coupled with refinement methods, and have contributed to the rapid proliferation of new structural information. With continual improvement in computer hardware and software for three-dimensional visualization of molecules (the crystallographer's version of 'virtual reality'), continuing rapid progress and evolution in this area are likely. The availability of computer graphics has also contributed greatly to the magnificent illustration of crystal structures, one of the factors that have thrust structural biology into many prominent roles in modern life and chemical sciences. Chapter 17.1 describes a graphics software product, *Coot*, now widely used by macromolecular crystallographers for visualization and real-space refinement of X-ray crystal structures. Chapter 17.2 surveys the field of computer visualization and animation of molecular structures, with a valuable historical perspective. Chapter 3.3 of *International Tables* Volume B is a useful reference for basics of computer-graphics visualization of molecules.

As in other areas of crystallography, refinement methods are used to obtain the most complete and precise structural information from macromolecular crystallographic data. The often-limited resolution and other factors lead to underdetermination of structural parameters relative to small-molecule crystal structures. In addition to X-ray intensity observations, macromolecular refinement incorporates observations about the normal stereochemistry of molecules, thereby improving the data-to-parameter ratio. Whereas incorporation of geometrical restraints and constraints in macromolecular refinement was initially implemented about 30 years ago, it is now generally a publication prerequisite that this methodology be used in structure refinement. Basic principles of crystallographic refinement, including least-squares minimization, constrained refinement and restrained refinement, are described in Chapter 18.1. Simulated-annealing methods, discussed in Chapter 18.2, can accelerate convergence to a refined structure, and are now widely used in refining macromolecular crystal structures. Structure quality and target parameters for stereochemical constraints and restraints are discussed in Chapter 18.3. High-resolution refinement of macromolecular structures, including handling of hydrogen-atom positions, is discussed in Chapter 18.4. Estimation of coordinate error in structure refinement is discussed in Chapter 18.5. Chapters 18.6 to 18.11 summarize computer programs and packages in common use for macromolecular structure determination, refinement and analysis. These program systems include the *Crystallography & NMR System* (*CNS*, Chapter 18.6), the *TNT* refinement package (Chapter 18.7), *ARP* and *wARP* for automated model construction and refinement (Chapter 18.8), *SHELX* (Chapter 18.9), *PrimeX* and the *Schrödinger* computational chemistry suite (Chapter 18.10), and *PHENIX* (Chapter 18.11). This part concludes with a description of methods for structure determination in the presence of merohedral twinning (Chapter 18.12).

Part 19 is a collection of short reviews of alternative methods for studying macromolecular structure. Each can provide information complementary to that obtained from single-crystal X-ray diffraction methods. In fact, structural information obtained from nuclear magnetic resonance (NMR) spectroscopy or cryo-electron microscopy is now frequently used in initiating crystal structure solution *via* the molecular replacement method (Part 13). Neutron diffraction, discussed in Chapter 19.1, can be used to obtain high-precision information about hydrogen atoms in macromolecular structures. Electron-diffraction studies of thin crystals are yielding structural information to increasingly high resolution, often for problems where obtaining three-dimensional crystals is challenging (Chapter 19.2). Small-angle X-ray (Chapter 19.3) and neutron (Chapter 19.4) scattering studies can be used to obtain information about shape and electron-density contrast even in noncrystalline materials, and are especially informative in cases of large macromolecular assemblies (*e.g.* viruses and ribosomes). Fibre diffraction (Chapter 19.5) can be used to study the structure of fibrous biological molecules. The combination of electron microscopy and crystallography is helping to bridge molecular structure and multi-molecular ultrastructure in living cells. Cryo-electron microscopy and high-resolution electron microscopy have been applied to the study of detailed structures of noncrystalline molecules of increasing complexity (Chapter 19.6). NMR spectroscopy has become a central method in the determination of small and medium-sized protein structures (Chapter 19.7), and yields unique descriptions of molecular interactions and motion in solution. Continuing breakthroughs in NMR technology are expanding greatly the size range of structures that can be studied by NMR. Part 19 concludes with descriptions of software packages in common use for data processing and image reconstruction of single particles using electron-microscopy techniques. These include *SPIDER* and *SPIRE* (Chapter 19.8), *IMAGIC 4D* (Chapter 19.9) and *EMAN* (Chapter 19.10).

Energy and molecular-dynamics calculations already play an integral role in many approaches for refining macromolecular structures (Part 20). Simulation methods hold promise for greater understanding of the time course of macromolecular motion than can be obtained through painstaking experimental approaches. However, experimental structures are still the starting point for simulation methods, and the quality of simulations is judged relative to experimental observables. Chapters 20.1 and 20.2 present complementary surveys of the current field of energy and molecular-dynamics calculations.

Structure validation (Part 21) is an important part of macromolecular crystal structure determination. Owing in part to the low data-to-parameter ratio and to problems of model phase bias, it can be difficult to correct misinterpretations of structure that can occur at many stages of structure determination. Chapters 21.1, 21.2 and 21.3 present approaches to structure validation using a range of reference information about macromolecular structure, in addition to observed diffraction intensities. Structure-validation methods are especially important in cases where unusual or highly unexpected features are found in a new structure. Software packages widely used for structure validation, as well as the detection and correction of errors in structures, are described in Chapters 21.4 (*PROCHECK*), 21.5 (*KiNG* and kinemages) and 21.6 (*MolProbity*).

Part 22 presents a survey of many methods used in the analysis of macromolecular structure. Since macromolecular structures tend to be very complicated, it is essential to extract features, descriptions and representations that can simplify information in helpful ways. Calculations of molecular surface areas, volumes and solvent-accessible surface areas are discussed in Chapters 22.1 and 22.2. Useful generalizations relating surface areas buried at macromolecular interfaces and energies of association have emerged. Chapter 22.3 surveys the occurrence of hydrogen bonds in biological macromolecules. This treatment is complemented by a description of solvent structure in Chapter 23.5, below. Electrostatic interactions in proteins are described in Chapter 22.4. The Cambridge Structural Database is the most complete compendium of small-molecule structural data; its role in assessing macromolecular crystal structures is discussed in Chapter 22.5. Databases for the deposition of macromolecular coordinates are the subject of Part 24, below.

Part 23 surveys current knowledge of protein and nucleic acid architecture. Proliferation of structural data has created problems for classification schemes, which have been forced to co-evolve with new structural knowledge. Methods and computational algorithms for the structural classification of various protein tertiary folds and the recognition of separate structural domains are reviewed in Chapters 23.1 and 23.2. Systematic aspects of ligand binding to macromolecules are discussed in Chapter 23.3. A survey of nucleic acid structure, geometry and classification schemes is presented in Chapter 23.4. Solvent structure in macromolecular crystals is reviewed in Chapter 23.5. Halogen interactions with small organic molecules have been known for some time now, but halogen interactions with macromolecules have only begun to gain recognition in recent years. These interactions tend to be unfamiliar and, at times, non-intuitive to the structural biologist. For example, an iodine or bromine atom (*i.e.*, of a ligand) might interact with the carbonyl oxygen atom of a protein without electronic repulsion, provided the geometry of the interaction allows the halogen atom to serve as a Lewis acid instead of the expected Lewis base. Our current understanding of halogen–macromolecule interactions and the geometric parameters within which they operate are reviewed in Chapter 23.6.

With the proliferation of macromolecular structures, it has been necessary to establish databases as international resources for rapid access to, and archival of, primary structural data. The Protein Data Bank (PDB), which for almost thirty years was the depository for protein crystal (and later NMR) structures, was first established at Brookhaven National Laboratory in the United States and later expanded into the worldwide PDB (wwPDB). The wwPDB is now run by the Research Collaboratory for Structural Bioinformatics (RCSB PDB) based in the United States with additional deposition centres: the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI), now the Protein Data Bank in Europe (PDBe); the PDB Japan (PDBj); and the BioMagResBank (BMRB). Chapter 24.1 describes the organization and features of the wwPDB. The wwPDB permits ready access to the rapidly increasing store of macromolecular structural data *via* the Internet, as well as rapid correlation of structural data with other key life-sciences databases. The Nucleic Acid Database (NDB), containing nucleic acid structures with and without bound ligands and proteins, is presented in Chapter 24.2. The Biological Macromolecule Crystallization Database (BMCD), a repository for macromolecular crystallization data, is described in Chapter 24.3.

Chapter 25.1 provides a detailed history of the structure determination of lysozyme, the first enzyme crystal structure to be solved. This chapter serves as a guide to the process by which the lysozyme structure was solved. Although the specific methods used to determine macromolecular structures have changed, the overall process is similar and the reader should find this account entertaining as well as instructive.

references

International Tables for Crystallography (2012). Vol. F, Chapter 1.2, pp. 5–12.

Chapter 1.2. Historical background

M. G. Rossmann

1.2.1. Introduction

Crystallography ranks with astronomy as one of the oldest sciences. Crystals, in the form of precious stones and common minerals, have attractive properties on account of their symmetry and their refractive and reflective properties, which result in the undefinable quality called beauty. Natural philosophers have long pondered the unusual properties seen in the discontinuous surface morphologies of crystals. Hooke (1665) and Huygens (1690) came close to grasping the way repeating objects create discrete crystal faces with reproducible interfacial angles. The symmetry of mineral crystals was explored systematically in the 18th and 19th centuries by measuring the angles between crystal faces, leading to the classification into symmetry systems from triclinic to cubic and the construction of symmetry tables (Schoenflies, 1891; Hilton, 1903; Astbury *et al.*, 1935) – the predecessors of today's *International Tables*.

1.2.2. 1912 to the 1950s

It was not until the interpretation of the first X-ray diffraction experiments by Max von Laue and Peter Ewald in 1912 that it was possible to ascertain the size of the repeating unit in simple crystals. Lawrence Bragg, encouraged by his father, William Bragg, recast the Laue equations into the physically intuitive form now known as 'Bragg's law' (Bragg & Bragg, 1913). This set the stage for a large number of structure determinations of inorganic salts and metals. The discovery of simple structures (Bragg, 1913), such as that of NaCl, led to a good deal of acrimony, for crystals of such salts were shown to consist of a uniform distribution of positive and negative ions, rather than discrete molecules. These early structure determinations were based on trial and error (sometimes guided by the predictions of Pope and Barlow that were based on packing considerations) until a set of atomic positions could be found that satisfied the observed intensity distribution of the X-ray reflections. This gave rise to rather pessimistic estimates that structures with more than about four independent atomic parameters would not be solvable.

The gradual advance in X-ray crystallography required a systematic understanding and tabulation of space groups. Previously, only various aspects of three-dimensional symmetry operations appropriate for periodic lattices had been listed. Consequently, in 1935, the growing crystallographic community put together the first set of *Internationale Tabellen* (Hermann, 1935), containing diagrams and information on about 230 space groups. After World War II, these tables were enlarged and combined with Kathleen Lonsdale's structure-factor formulae (Lonsdale, 1936) in the form of *International Tables* Volume I (Henry & Lonsdale, 1952). Most recently, they have again been revised and extended in Volume A (Hahn, 2005).

Simple organic compounds started to be examined in the 1920s. Perhaps foremost among these is the structure of hexamethylbenzene by Kathleen Lonsdale (Lonsdale, 1928). She showed that, as had been expected, benzene had a planar hexagonal structure. Another notable achievement of crystallography was made by J. D. Bernal in the early 1930s. He was able to differentiate between a number of possible structures for steroids by studying their packing arrangements in different unit cells (Bernal, 1933). Bernal ('Sage') had an enormous impact on English crystallographers in the 1930s. His character was immortalized by the novelist C. P. Snow in his book *The Search* (Snow, 1934). By the mid-1930s, J. Monteath Robertson and I. Woodward had determined the structure of nickel phthalocyanine (Robertson, 1935) using the heavy-atom method. This was a major crystallographic success and perhaps the first time that a crystallographer had succeeded in solving a structure when little chemical information was available.

Another event which had a major impact was the determination of the absolute hand of the asymmetric carbon atom of sodium tartrate by Bijvoet (Bijvoet, 1949; Bijvoet *et al.*, 1951). By indexing the X-ray reflections with a right-handed system, he showed that the breakdown of Friedel's law in the presence of an anomalous scatterer was consistent with the asymmetric carbon atom having a hand in agreement with Fischer's convention. With that knowledge, together with the prior results of organic reaction analyses, the absolute hand of other asymmetric carbon atoms could be established. In particular, the absolute structure of naturally occurring amino acids and riboses was now determined.

Until the mid-1950s, most structure determinations were made using only projection data. This not only reduced the tremendous effort required for manual indexing and for making visual estimates of intensity measurements, but also reduced the calculation effort to almost manageable proportions in the absence of computing machines. However, the structure determination of penicillin (Crowfoot, 1948; Crowfoot *et al.*, 1949), carried out during World War II by Dorothy Hodgkin and Charles Bunn, employed some three-dimensional data. A further major achievement was the solution of the three-dimensional structure of vitamin B_{12} by Dorothy Hodgkin and her colleagues (Hodgkin *et al.*, 1957) in the 1950s. They first used a cobalt atom as a heavy atom on a vitamin B_{12} fragment and were able to recognize the 'corrin' ring structure. The remainder of the B_{12} structure was determined by an extraordinary collaboration between Dorothy Hodgkin in Oxford and Kenneth Trueblood at UCLA in Los Angeles. While Dorothy's group did the data collection and interpretation, Ken's group performed the computing on the very early electronic Standard Western Automatic Computer (SWAC). Additional help was made available by the parallel work of J. G. White at Princeton University in New Jersey. This was at a time before the internet, before e-mail, before usable transatlantic telephones and before jet travel. Transatlantic, propeller-driven air connections had started to operate only a few years earlier.

Many technical advances were made in the 1930s that contributed to the rapidly increasing achievements of crystallography. W. H. Bragg had earlier suggested (Bragg, 1915) the use of Fourier methods to analyse the periodic electron-density distribution in crystals, and this was utilized by his son, W. L. Bragg (Bragg, 1929*a*,*b*). The relationship between a Fourier synthesis and a Fourier analysis demonstrated that the central problem in structural crystallography was in the phase. Compu-

Copyright © 2012 International Union of Crystallography

tational devices to help plot this distribution were invented by Arnold Beevers and Henry Lipson in the form of their 'Beevers–Lipson strips' (Beevers & Lipson, 1934) and by J. Monteath Robertson with his 'Robertson sorting board' (Robertson, 1936). These devices were later supplemented by the XRAC electronic analogue machine of Ray Pepinsky (Pepinsky, 1947) and mechanical analogue machines (McLachlan & Champaygne, 1946; Lipson & Cochran, 1953) until electronic digital computers came into use during the mid-1950s.

A. Lindo Patterson, inspired by his visit to England in the 1930s where he met Lawrence Bragg, Kathleen Lonsdale and J. Monteath Robertson, showed how to use F^2 Fourier syntheses for structure determinations (Patterson, 1934, 1935). When the 'Patterson' synthesis was combined with the heavy-atom method, and (later) with electronic computers, it transformed analytical organic chemistry. No longer was it necessary for teams of chemists to labour for decades on the structure determination of natural products. Instead, a single crystallographer could solve such a structure in a period of months.

Improvements in data-collection devices have also had a major impact. Until the mid-1950s, the most common method of measuring intensities was by visual comparison of reflection 'spots' on films with a standard scale. However, the use of counters (used, for instance, by Bragg in 1912) was gradually automated and became the preferred technique in the 1960s. In addition, semi-automatic methods of measuring the optical densities along reciprocal lines on precession photographs were used extensively for early protein-structure determinations in the 1950s and 1960s.

1.2.3. The first investigations of biological macromolecules

Leeds, in the county of Yorkshire, was one of the centres of England's textile industry and home to a small research institute established to investigate the properties of natural fibres. W. T. Astbury became a member of this institute after learning about X-ray diffraction from single crystals in Bragg's laboratory. He investigated the diffraction of X-rays by wool, silk, keratin and other natural fibrous proteins. He showed that the resultant patterns could be roughly classified into two classes, α and β, and that on stretching some, for example, wool, the pattern is converted from α to β (Astbury, 1933).

Purification techniques for globular proteins were also being developed in the 1920s and 1930s, permitting J. B. Sumner at Cornell University to crystallize the first enzyme, namely urease, in 1926. Not much later, in Cambridge, J. D. Bernal and his student, Dorothy Crowfoot (Hodgkin), investigated crystals of pepsin. The resultant 1934 paper in *Nature* (Bernal & Crowfoot, 1934) is quite remarkable because of its speed of publication and because of the authors' extraordinary insight. The crystals of pepsin were found to deteriorate quickly in air when taken out of their crystallization solution and, therefore, had to be contained in a sealed capillary tube for all X-ray experiments. This form of protein-crystal mounting remained in vogue until the 1990s when crystal-freezing techniques were introduced. But, most importantly, it was recognized that the pepsin diffraction pattern implied that the protein molecules have a unique structure and that these crystals would be a vehicle for the determination of that structure to atomic resolution. This understanding of protein structure occurred at a time when proteins were widely thought to form heterogeneous micelles, a concept which persisted another 20 years until Sanger was able to determine the unique amino-acid sequences of the two chains in an insulin molecule (Sanger & Tuppy, 1951; Sanger & Thompson, 1953*a*,*b*).

Soon after Bernal and Hodgkin photographed an X-ray diffraction pattern of pepsin, Max Perutz started his historic investigation of haemoglobin.[1] Such investigations were, however, thought to be without hope of any success by most of the contemporary crystallographers, who avoided crystals that did not have a short (less than 4.5 Å) axis for projecting resolved atoms. Nevertheless, Perutz computed Patterson functions that suggested haemoglobin contained parallel α-keratin-like bundles of rods (Boyes-Watson *et al.*, 1947; Perutz, 1949). Perutz was correct about the α-keratin-like rods, but not about these being parallel.

In Pasadena, Pauling (Pauling & Corey, 1951; Pauling *et al.*, 1951) was building helical polypeptide models to explain Astbury's α patterns and perhaps to understand the helical structures in globular proteins, such as haemoglobin. Pauling, using his knowledge of the structure of amino acids and peptide bonds, was forced to the conclusion that there need not be an integral number of amino-acid residues per helical turn. He therefore suggested that the 'α-helix', with 3.6 residues per turn, would roughly explain Astbury's α pattern and that his proposed 'β-sheet' structure should be related to Astbury's β pattern. Perutz saw that an α-helical structure should give rise to a strong 1.5 Å-spacing reflection as a consequence of the rise per residue in an α-helix (Perutz, 1951*a*,*b*). Demonstration of this reflection in horse hair, then in fibres of polybenzyl-L-glutamate, in muscle (with Hugh Huxley) and finally in haemoglobin crystals showed that Pauling's proposed α-helix really existed in haemoglobin and presumably also in other globular proteins. Confirmation of helix-like structures came with the observation of cylindrical rods in the 6 Å-resolution structure of myoglobin in 1957 (Kendrew *et al.*, 1958) and eventually at atomic resolution with the 2 Å myoglobin structure in 1959 (Kendrew *et al.*, 1960). The first atomic resolution confirmation of Pauling's β structure did not come until 1966 with the structure determination of hen egg-white lysozyme (Blake, Mair *et al.*, 1967).

Although the stimulus for the Cochran *et al.* (1952) analysis of diffraction from helical structures came from Perutz's studies of helices in polybenzyl-L-glutamate and their presence in haemoglobin, the impact on the structure determination of nucleic acids was even more significant. The events leading to the discovery of the double-helical structure of DNA have been well chronicled (Watson, 1968; Olby, 1974; Judson, 1979). The resultant science, often known exclusively as molecular biology, has created a whole new industry. Furthermore, the molecular-modelling techniques used by Pauling in predicting the structure of α-helices and β-sheets and by Crick and Watson in determining the structure of DNA had a major effect on more traditional crystallography and the structure determinations of fibrous proteins, nucleic acids and polysaccharides.

Another major early result of profound biological significance was the demonstration by Bernal and Fankuchen in the 1930s (Bernal & Fankuchen, 1941) that tobacco mosaic virus (TMV) had a rod-like structure. This was the first occasion where it was possible to obtain a definite idea of the architecture of a virus.

---

[1] Perutz writes, 'I started X-ray work on haemoglobin in October 1937 and Bragg became Cavendish Professor in October 1938. Bernal was my PhD supervisor in 1937, but he had nothing to do with my choice of haemoglobin. I began this work at the suggestion of Haurowitz, the husband of my cousin Gina Perutz, who was then in Prague. The first paper on X-ray diffraction from haemoglobin (and chymotrypsin) was Bernal, Fankuchen & Perutz (Bernal *et al.*, 1938). I did the experimental work, (and) Bernal showed me how to interpret the X-ray pictures.'

Many of the biological properties of TMV had been explored by Wendell Stanley working at the Rockefeller Institute in New York. He had also been able to obtain a large amount of purified virus. Although it was not possible to crystallize this virus, it was possible to obtain a diffraction pattern of the virus in a viscous solution which had been agitated to cause alignment of the virus particles. This led Jim Watson (Watson, 1954) to a simple helical structure of protein subunits. Eventually, after continuing studies by Aaron Klug, Rosalind Franklin, Ken Holmes and others, the structure was determined at atomic resolution (Holmes *et al.*, 1975), in which the helical strand of RNA was protected by the helical array of protein subunits.

1.2.4. Globular proteins in the 1950s

In 1936, Max Perutz had joined Sir Lawrence Bragg in Cambridge. Inspired in part by Keilin (Perutz, 1997), Perutz started to study crystalline haemoglobin. This work was interrupted by World War II, but once the war was over Perutz tenaciously developed a series of highly ingenious techniques. All of these procedures have their counterparts in modern 'protein crystallography', although few today recognize their real origin.

The first of these methods was the use of 'shrinkage' stages (Perutz, 1946; Bragg & Perutz, 1952). It had been noted by Bernal and Crowfoot (Hodgkin) in their study of pepsin that crystals of proteins deteriorate on exposure to air. Perutz examined crystals of horse haemoglobin after they were air-dried for short periods of time and then sealed in capillaries. He found that there were at least seven consecutive discrete shrinkage stages of the unit cell. He realized that each shrinkage stage permitted the sampling of the molecular transform at successive positions, thus permitting him to map the variation of the continuous transform. As he examined only the centric ($h0l$) reflections of the monoclinic crystals, he could observe when the sign changed from 0 to π in the centric projection (Fig. 1.2.4.1). Thus, he was able to determine the phases (signs) of the central part of the ($h0l$) reciprocal lattice. This technique is essentially identical to the use of diffraction data from different unit cells for averaging electron density in the 'modern' molecular replacement method. In the haemoglobin case, Patterson projections had shown that the molecules maintained their orientation relative to the a axis as the crystals shrank, but in the more general molecular replacement case, it is necessary to determine the relative orientations of the molecules in each cell.

The second of Perutz's techniques depended on observing changes in the intensities of low-order reflections when the concentration of the dissolved salts (*e.g.* Cs_2SO_4) in the solution between the crystallized molecules was altered (Boyes-Watson *et al.*, 1947; Perutz, 1954). The differences in structure amplitude, taken together with the previously determined signs, could then map out the parts of the crystal unit cell occupied by the haemoglobin molecule. In many respects, this procedure has its equivalent in 'solvent flattening' used extensively in 'modern' protein crystallography.

The third of Perutz's innovations was the isomorphous replacement method (Green *et al.*, 1954). The origin of the isomorphous replacement method goes back to the beginnings of X-ray crystallography when Bragg compared the diffracted intensities from crystals of NaCl and KCl. J. Monteath Robertson explored the procedure a little further in his studies of phthalocyanines. Perutz used a well known fact that dyes could be diffused into protein crystals, and, hence, heavy-atom compounds might also diffuse into and bind to specific residues in the protein. Nevertheless, the sceptics questioned whether even the heaviest atoms could make a measurable difference to the X-ray diffraction pattern of a protein.[2] Perutz therefore developed an instrument which quantitatively recorded the blackening caused by the reflected X-ray beam on a film. He also showed that the effect of specifically bound atoms could be observed visually on a film record of a diffraction pattern. In 1953, this resulted in a complete sign determination of the ($h0l$) horse haemoglobin structure amplitudes (Green *et al.*, 1954). However, not surprisingly, the projection of the molecule was not very interesting, making it necessary to extend the procedure to noncentric, three-dimensional data. It took another five years to determine the first globular protein structure to near atomic resolution.

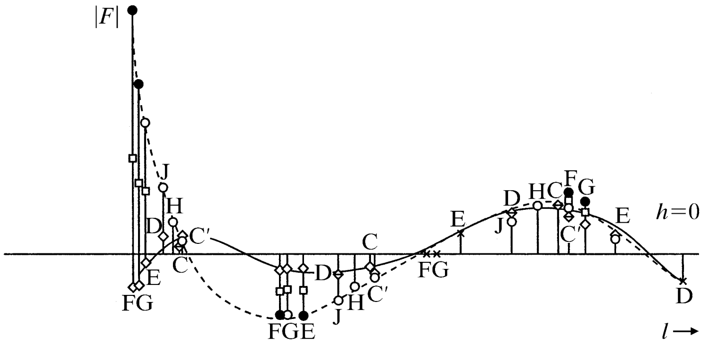


Figure 1.2.4.1
Change of structure amplitude for horse haemoglobin as a function of salt concentration in the suspension medium of the low-order $h0l$ reflections at various lattice shrinkage stages (C, C′, D, E, F, G, H, J). Reprinted with permission from Perutz (1954). Copyright (1954) Royal Society of London.

In 1950, David Harker was awarded one million US dollars to study the structure of proteins. He worked first at the Brooklyn Polytechnic Institute in New York and later at the Roswell Park Cancer Institute in Buffalo, New York. He proposed to solve the structure of proteins on the assumption that they consisted of 'globs' which he could treat as single atoms; therefore, he could solve the structure by using his inequalities (Harker & Kasper, 1947), *i.e.*, by direct methods. He was aware of the need to use three-dimensional data, which meant a full phase determination, rather than the sign determination of two-dimensional projection data on which Perutz had concentrated. Harker therefore decided to develop automatic diffractometers, as opposed to the film methods being used at Cambridge. In 1956, he published a procedure for plotting the isomorphous data of each reflection in a simple graphical manner that allowed an easy determination of its phase (Harker, 1956). Unfortunately, the error associated with the data tended to create a lot of uncertainty.

In the first systematic phase determination of a protein, namely that of myoglobin, phase estimates were made for about 400 reflections. In order to remove subjectivity, independent estimates were made by Kendrew and Bragg by visual inspection of the Harker diagram for each reflection. These were later compared before computing an electron-density map. This process was put onto a more objective basis by calculating phase probabilities, as described by Blow & Crick (1959) and Dickerson *et al.* (1961).

---

[2] Perutz writes, 'I measured the absolute intensity of reflexions from haemoglobin which turned out to be weaker than predicted by Wilson's statistics. This made me realise that about 99% of the scattering contributions of the light atoms are extinguished by interference and that, by contrast, the electrons of a heavy atom, being concentrated at a point, would scatter in phase and therefore make a measurable difference to the structure amplitudes.'

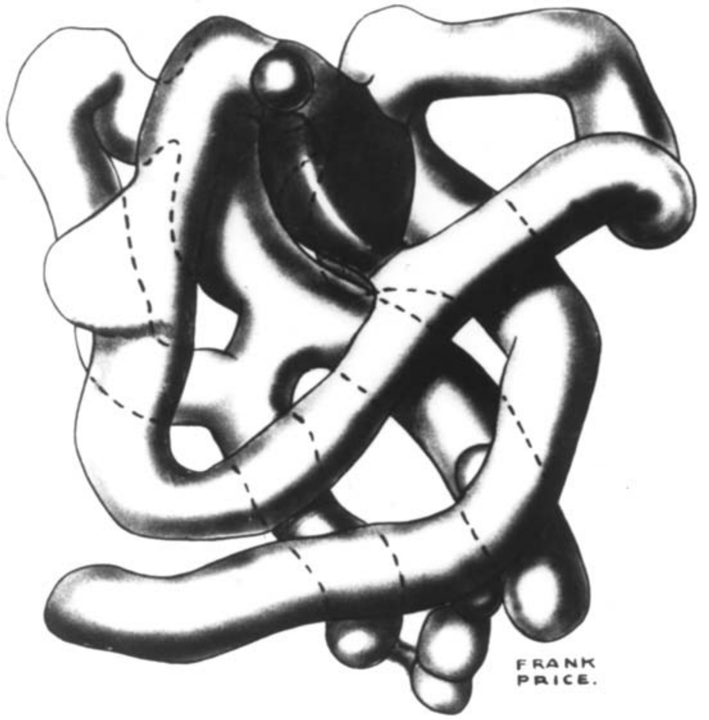


Figure 1.2.5.1
A model of the myoglobin molecule at 6 Å resolution. Reprinted with permission from Bodo *et al.* (1959). Copyright (1959) Royal Society of London.

One problem with the isomorphous replacement method was the determination of accurate parameters that described the heavy-atom replacements. Centric projections were a means of directly determining the coordinates, but no satisfactory method was available to determine the relative positions of atoms in different derivatives when there were no centric projections. In particular, it was necessary to establish the relative y coordinates for the heavy-atom sites in the various derivatives of monoclinic myoglobin and in monoclinic horse haemoglobin. Perutz (1956) and Bragg (1958) had each proposed solutions to this problem, but these were not entirely satisfactory. Consequently, it was necessary to average the results of different methods to determine the 6 Å phases for myoglobin. However, this problem was solved satisfactorily in the structure determination of haemoglobin by using an $(F_{H1} - F_{H2})^2$ Patterson-like synthesis in which the vectors between atoms in the two heavy-atom compounds, H1 and H2, produce negative peaks (Rossmann, 1960). This technique also gave rise to the first least-squares refinement procedure to determine the parameters that define each heavy atom.

Perutz used punched cards to compute the first three-dimensional Patterson map of haemoglobin. This was a tremendous computational undertaking. However, the first digital electronic computers started to appear in the early to mid-1950s. The EDSAC1 and EDSAC2 machines were built in the Mathematical Laboratory of Cambridge University. EDSAC1 was used by John Kendrew for the 6 Å-resolution map of myoglobin (Bluhm *et al.*, 1958). EDSAC2 came online in 1958 and was the computer on which all the calculations were made for the 5.5 Å map of haemoglobin (Cullis *et al.*, 1962) and the 2.0 Å map of myoglobin. It was the tool on which many of the now well established crystallographic techniques were initially developed. By about 1960, the home-built, one-of-a-kind machines were starting to be replaced by commercial machines. Large mainframe IBM computers (704, 709 *etc.*), together with FORTRAN as a symbolic language, became available.

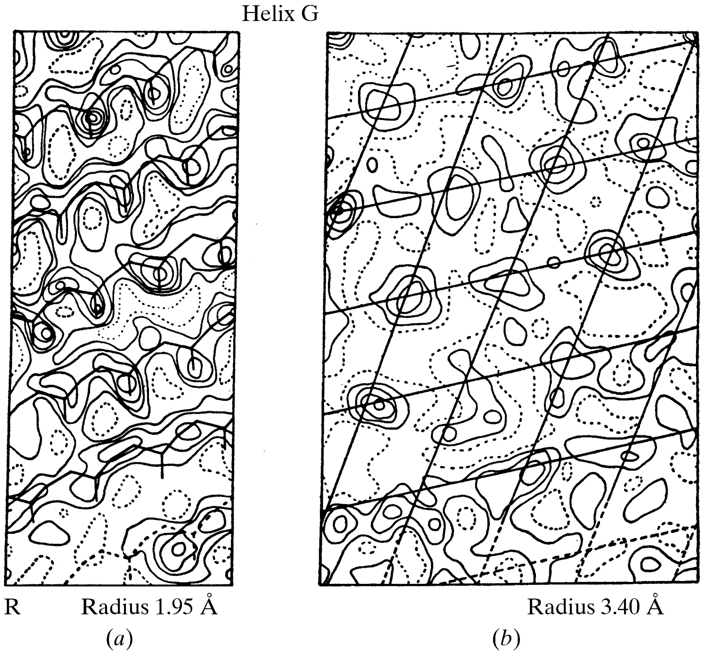


Figure 1.2.5.2
Cylindrical sections through a helical segment of a myoglobin polypeptide chain. (*a*) The density in a cylindrical mantle of 1.95 Å radius, corresponding to the mean radius of the main-chain atoms in an α-helix. The calculated atomic positions of the α-helix are superimposed and roughly correspond to the density peaks. (*b*) The density at the radius of the β-carbon atoms; the positions of the β-carbon atoms calculated for a right-handed α-helix are marked by the superimposed grid (Kendrew & Watson, unpublished). Reprinted with permission from Perutz (1962). Copyright (1962) Elsevier Publishing Co.

1.2.5. The first protein structures (1957 to the 1970s)

By the time three-dimensional structures of proteins were being solved, Linderström-Lang (Linderström-Lang & Schellman, 1959) had introduced the concepts of 'primary', 'secondary' and 'tertiary' structures, providing a basis for the interpretation of electron-density maps. The first three-dimensional protein structure to be solved was that of myoglobin at 6 Å resolution (Fig. 1.2.5.1) in 1957 (Kendrew *et al.*, 1958). It clearly showed sausage-like features which were assumed to be α-helices. The iron-containing haem group was identified as a somewhat larger electron-density feature. The structure determination of haemoglobin at 5.5 Å resolution in 1959 (Cullis *et al.*, 1962) showed that each of its two independent chains, α and β, had a fold similar to that of myoglobin and, thus, suggested a divergent evolutionary process for oxygen transport molecules. These first protein structures were mostly helical, features that could be recognized readily at low resolution. Had the first structures been of mostly β structure, as is the case for pepsin or chymotrypsin, history might have been different.

The absolute hand of the haemoglobin structure was determined using anomalous dispersion (Cullis *et al.*, 1962) in a manner similar to that used by Bijvoet. This was confirmed almost immediately when a 2 Å-resolution map of myoglobin was calculated in 1959 (Kendrew *et al.*, 1960). By plotting the electron density of the α-helices on cylindrical sections (Fig. 1.2.5.2), it was possible to see not only that the Pauling prediction of the α-helix structure was accurately obeyed, but also that the C_β atoms were consistent with *laevo* amino acids and that all eight helices were right-handed on account of the steric hindrance that would occur between the C_β atom and carbonyl oxygen in left-handed helices.

The first enzyme structure to be solved was that of lysozyme in 1965 (Blake *et al.*, 1965), following a gap of six years after the excitement caused by the discovery of the globin structures. Diffusion of substrates into crystals of lysozyme showed how substrates bound to the enzyme (Blake, Johnson *et al.*, 1967), which in turn suggested a catalytic mechanism and identified the essential catalytic residues.

From 1965 onwards, the rate of protein-structure determinations gradually increased to about one a year: carboxypeptidase (Reeke *et al.*, 1967), chymotrypsin (Matthews *et al.*, 1967), ribonuclease (Kartha *et al.*, 1967; Wyckoff *et al.*, 1967), papain (Drenth *et al.*, 1968), insulin (Adams *et al.*, 1969), lactate dehydrogenase (Adams *et al.*, 1970) and cytochrome *c* (Dickerson *et al.*, 1971) were early examples. Every new structure was a major event. These structures laid the foundation for structural biology. From a crystallographic point of view, Drenth's structure determination of papain was particularly significant in that he was able to show an amino-acid sequencing error where 13 residues had to be inserted between Phe28 and Arg31, and he showed that a 38-residue peptide that had been assigned to position 138 to 176 needed to be transposed to a position between the extra 13 residues and Arg31.

The structures of the globins had suggested that proteins with similar functions were likely to have evolved from a common precursor and, hence, that there might be a limited number of protein-folding motifs. Comparison of the active centres of chymotrypsin and subtilisin showed that convergent evolutionary pathways could exist (Drenth *et al.*, 1972; Kraut *et al.*, 1972).

The variety of structures that were being studied increased rapidly. The first tRNA structures were determined in the 1960s (Kim *et al.*, 1973; Robertus *et al.*, 1974), the first spherical virus structure was published in 1978 (Harrison *et al.*, 1978) and the photoreaction centre membrane protein structure appeared in 1985 (Deisenhofer *et al.*, 1985). The rate of new structure determinations has continued to increase exponentially. In 1996, about one new structure was published every day. Partly in anticipation and partly to assure the availability of results, the Brookhaven Protein Data Bank (PDB) was brought to life at the 1971 Cold Spring Harbor Meeting (H. Berman & J. L. Sussman, private communication). Initially, it was difficult to persuade authors to submit their coordinates, but gradually this situation changed to one where most journals require coordinate submission to the PDB, resulting in today's access to structural results *via* the World Wide Web.

The growth of structural information permitted generalizations, such as that β-sheets have a left-handed twist when going from one strand to the next (Chothia, 1973) and that 'cross-over' β-α-β turns were almost invariably right-handed (Richardson, 1977). These observations and the growth of the PDB have opened up a new field of science. Among the many important results that have emerged from this wealth of data is a careful measurement of the main-chain dihedral angles, confirming the predictions of Ramachandran (Ramachandran & Sasisekharan, 1968), and of side-chain rotamers (Ponder & Richards, 1987). Furthermore, it is now possible to determine whether the folds of domains in a new structure relate to any previous results quite conveniently (Murzin *et al.*, 1995; Holm & Sander, 1997).

1.2.6. Technological developments (1958 to the 1980s)

In the early 1960s, there were very few who had experience in solving a protein structure. Thus, almost a decade passed after the success with the globins before there was a noticeable surge of new structure reports. In the meantime, there were persistent attempts to find alternative methods to determine protein structure.

Blow & Rossmann (1961) demonstrated the power of the single isomorphous replacement method. While previously it had been thought that it was necessary to have at least two heavy-atom compounds, if not many more, they showed that a good representation of the structure of haemoglobin could have been made by using only one good derivative. There were also early attempts at exploiting anomalous dispersion for phase determination. Rossmann (1961) showed that anomalous differences could be used to calculate a 'Bijvoet Patterson' from which the sites of the anomalous scatterers (and, hence, heavy-atom sites) could be determined. Blow & Rossmann (1961), North (1965) and Matthews (1966) used anomalous-dispersion data to help in phase determination. Hendrickson stimulated further interest by using Cu $K\alpha$ radiation and employing the anomalous effect of sulfur atoms in cysteines to solve the entire structure of the crambin molecule (Hendrickson & Teeter, 1981). With today's availability of synchrotrons, and hence the ability to tune to absorption edges, these early attempts to utilize anomalous data have been vastly extended to the powerful multiple-wavelength anomalous dispersion (MAD) method (Hendrickson, 1991). More recently, the generality of the MAD technique has been greatly expanded by using proteins in which methionine residues have been replaced by selenomethionine, thus introducing selenium atoms as anomalous scatterers.

Another advance was the introduction of the 'molecular replacement' technique (Rossmann, 1972). The inspiration for this method arose out of the observation that many larger proteins (*e.g.* haemoglobin) are oligomers of identical subunits and that many proteins can crystallize in numerous different forms. Rossmann & Blow (1962) recognized that an obvious application of the technique would be to viruses with their icosahedral symmetry. They pointed out that the symmetry of the biological oligomer can often be, and sometimes must be, 'noncrystallographic' or 'local', as opposed to being true for the whole infinite crystal lattice. Although the conservation of folds had become apparent in the study of the globins and a little later in the study of dehydrogenases (Rossmann *et al.*, 1974), in the 1960s the early development of the molecular replacement technique was aimed primarily at *ab initio* phase determination (Rossmann & Blow, 1963; Main & Rossmann, 1966; Crowther, 1969). It was only in the 1970s, when more structures became available, that it was possible to use the technique to solve homologous structures with suitable search models. Initially, there was a good deal of resistance to the use of the molecular replacement technique. Results from the rotation function were often treated with scepticism, the translation problem was thought to have no definitive answer, and there were excellent reasons to consider that phasing was impossible except for centric reflections (Rossmann, 1972). It took 25 years before the full power of all aspects of the molecular replacement technique was fully utilized and accepted (Rossmann *et al.*, 1985).

The first real success of the rotation function was in finding the rotational relationship between the two independent insulin monomers in the *P*3 unit cell (Dodson *et al.*, 1966). Crowther produced the fast rotation function, which reduced the computational times to manageable proportions (Crowther, 1972). Crowther (1969) and Main & Rossmann (1966) were able to formulate the problem of phasing in the presence of noncrystallographic symmetry in terms of a simple set of simultaneous

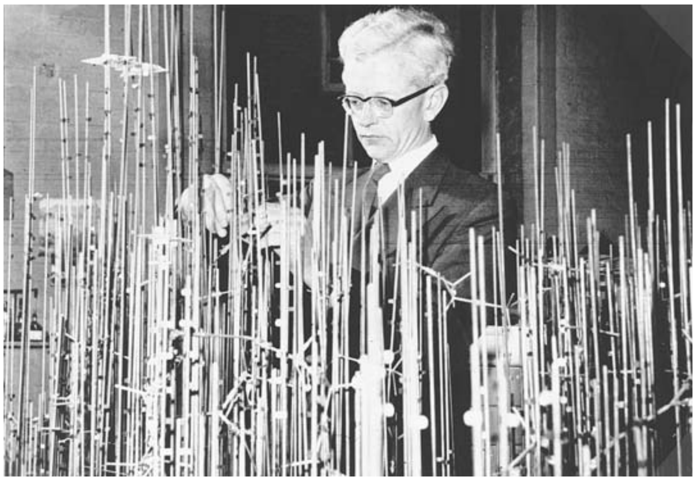


Figure 1.2.6.1
The 2 Å-resolution map of sperm-whale myoglobin was represented by coloured Meccano-set clips on a forest of vertical rods. Each clip was at a grid point. The colour of the clip indicated the height of the electron density. The density was interpreted in terms of 'Corey–Pauling' models on a scale of 5 cm = 1 Å. Pictured is John Kendrew. (This figure was provided by M. F. Perutz.)

complex equations. However, real advances came with applying the conditions of noncrystallographic symmetry in real space, which was the key to the solution of glyceraldehyde-3-phosphate dehydrogenase (Buehner *et al.*, 1974), tobacco mosaic virus disk protein (Bloomer *et al.*, 1978) and other structures, aided by Gerard Bricogne's program for electron-density averaging (Bricogne, 1976), which became a standard of excellence.

No account of the early history of protein crystallography is complete without a mention of ways of representing electron density. The 2 Å map of myoglobin was interpreted by building a model (on a scale of 5 cm to 1 Å with parts designed by Corey and Pauling at the California Institute of Technology) into a forest of vertical rods decorated with coloured clips at each grid point, representing the height of the electron density (Fig. 1.2.6.1). Later structures, such as those of lysozyme and carboxypeptidase, were built with 'Kendrew' models (2 cm to 1 Å) based on electron-density maps displayed as stacks of large Plexiglas sheets. A major advance came with Fred Richards' invention of the optical comparator (a 'Richards box' or 'Fred's folly') in which the model was optically superimposed onto the electron density by reflection of the model in a half-silvered mirror (Richards, 1968). This allowed for convenient fitting of model parts and accurate placement of atoms within the electron density. The Richards box was the forerunner of today's computer graphics, originally referred to as an 'electronic Richards box'. The development of computer graphics for model building was initially met with reservation, but fortunately those involved in these developments persevered. Various programs became available for model building in a computer, but the undoubted champion of this technology was *FRODO*, written by Alwyn Jones (Jones, 1978).

1.2.7. Meetings

The birth of protein crystallography in the 1950s coincided with the start of the jet age, making attendance at international meetings far easier. Indeed, the number and variety of meetings have proliferated as much as the number and variety of structures determined. A critical first for protein crystallography was a meeting held at the Hirschegg ski resort in Austria in 1966. This was organized by Max Perutz (Cambridge) and Walter Hoppe (München). About 40 scientists from around the world attended, as well as a strong representation of students (including Robert Huber) from the Münich laboratory. The first Hirschegg meeting occurred just after the structure determination of lysozyme, which helped lift the cloud of pessimism after the long wait for a new structure since the structures of the globins had been solved in the 1950s. The meeting was very much a family affair where most attendees stayed an extra few days for additional skiing. The times were more relaxed in comparison with those of today's jet-setting scientists flying directly from synchrotron to international meeting, making ever more numerous important discoveries. A second Hirschegg meeting occurred two years later, but this time the number of participants had doubled. By 1970, the meeting had to be transferred to the village of Alpbach, which had more accommodation; however, most of the participants still knew each other.

Another set of international meetings (or schools, as the Italians preferred to call them) was initiated by the Italian crystallographers in 1976 at Erice, a medieval hilltop town in Sicily. These meetings have since been repeated every six years. The local organizer was Lodovico Riva di Sanseverino, whose vivacious sensitivity instilled a feeling of international fellowship into the rapidly growing number of structural biologists. The first meeting lasted two whole weeks, a length of time that would no longer be acceptable in today's hectic, competitive atmosphere.

It took time for the staid organizers of the IUCr triennial congress to recognize the significance of macromolecular structure. Thus, for many years, the IUCr meetings were poorly attended by structural biologists. However, recent meetings have changed, with biological topics representing about half of all activities. Nevertheless, the size of these meetings and their lack of focus have led to numerous large and small specialized meetings, from small Gordon Conferences and East and West Coast crystallography meetings in the USA to huge international congresses in virology, biochemistry and other sciences.

The publication of this volume by the International Union of Crystallography, the first volume of *International Tables* devoted to macromolecular crystallography, strongly attests to the increasing importance of this vital area of science.

I am gratefully indebted to Sharon Wilder, who has done much of the background checking that was required to write this chapter. Furthermore, she has been my permanent and faithful helper during a time that is, in part, covered in the review. I am additionally indebted to Max Perutz and David Davies, both of whom read the manuscript very carefully, making it possible to add a few personal accounts. I also thank the National Institutes of Health and the National Science Foundation for generous financial support.

References

Adams, M. J., Blundell, T. L., Dodson, E. J., Dodson, G. G., Vijayan, M., Baker, E. N., Harding, M. M., Hodgkin, D. C., Rimmer, B. & Sheat, S. (1969). *Structure of rhombohedral 2 zinc insulin crystals. Nature (London)*, **224**, 491–495.

Adams, M. J., Ford, G. C., Koekoek, R., Lentz, P. J. Jr, McPherson, A. Jr, Rossmann, M. G., Smiley, I. E., Schevitz, R. W. & Wonacott, A. J. (1970). *Structure of lactate dehydrogenase at 2.8 Å resolution. Nature (London)*, **227**, 1098–1103.

Astbury, W. T. (1933). *Fundamentals of Fiber Structure.* Oxford University Press.

Astbury, W. T., Mauguin, C., Hermann, C., Niggli, P., Brandenberger, E. & Lonsdale-Yardley, K. (1935). *Space groups*. In *Internationale*

Tabellen zur Bestimmung von Kristallstrukturen, edited by C. Hermann, Vol. 1, pp. 84–87. Berlin: Begrüder Borntraeger.

Beevers, C. A. & Lipson, H. (1934). *A rapid method for the summation of a two-dimensional Fourier series. Philos. Mag.* **17**, 855–859.

Bernal, J. D. (1933). *Comments. Chem. Ind.* **52**, 288.

Bernal, J. D. & Crowfoot, D. (1934). *X-ray photographs of crystalline pepsin. Nature (London)*, **133**, 794–795.

Bernal, J. D. & Fankuchen, I. (1941). *X-ray and crystallographic studies of plant virus preparations. J. Gen. Physiol.* **25**, 111–165.

Bernal, J. D., Fankuchen, I. & Perutz, M. (1938). *An X-ray study of chymotrypsin and haemoglobin. Nature (London)*, **141**, 523–524.

Bijvoet, J. M. (1949). *K. Ned. Akad. Wet.* **52**, 313.

Bijvoet, J. M., Peerdeman, A. F. & van Bommel, A. J. (1951). *Determination of the absolute configuration of optically active compounds by means of X-rays. Nature (London)*, **168**, 271–272.

Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Crystallographic studies of the activity of hen egg-white lysozyme. Proc. R. Soc. London Ser. B*, **167**, 378–388.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *On the conformation of the hen egg-white lysozyme molecule. Proc. R. Soc. London. Ser. B*, **167**, 365–377.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. (1978). *Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. Nature (London)*, **276**, 362–368.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method. Acta Cryst.* **14**, 1195–1202.

Bluhm, M. M., Bodo, G., Dintzis, H. M. & Kendrew, J. C. (1958). *The crystal structure of myoglobin. IV. A Fourier projection of sperm-whale myoglobin by the method of isomorphous replacement. Proc. R. Soc. London Ser. A*, **246**, 369–389.

Bodo, G., Dintzis, H. M., Kendrew, J. C. & Wyckoff, H. W. (1959). *The crystal structure of myoglobin. V. A low-resolution three-dimensional Fourier synthesis of sperm-whale myoglobin crystals. Proc. R. Soc. London Ser. A*, **253**, 70–102.

Boyes-Watson, J., Davidson, E. & Perutz, M. F. (1947). *An X-ray study of horse methaemoglobin. I. Proc. R. Soc. London Ser. A*, **191**, 83–132.

Bragg, L. & Perutz, M. F. (1952). *The structure of haemoglobin. Proc. R. Soc. London Ser. A*, **213**, 425–435.

Bragg, W. H. (1915). *X-rays and crystal structure. Philos. Trans. R. Soc. London Ser. A*, **215**, 253–274.

Bragg, W. H. & Bragg, W. L. (1913). *The reflection of X-rays in crystals. Proc. R. Soc. London Ser. A*, **88**, 428–438.

Bragg, W. L. (1913). *The structure of some crystals as indicated by their diffraction of X-rays. Proc. R. Soc. London Ser. A*, **89**, 248–277.

Bragg, W. L. (1929*a*). *The determination of parameters in crystal structures by means of Fourier series. Proc. R. Soc. London Ser. A*, **123**, 537–559.

Bragg, W. L. (1929*b*). *An optical method of representing the results of X-ray analysis. Z. Kristallogr.* **70**, 475–492.

Bragg, W. L. (1958). *The determination of the coordinates of heavy atoms in protein crystals. Acta Cryst.* **11**, 70–75.

Bricogne, G. (1976). *Methods and programs for direct-space exploitation of geometric redundancies. Acta Cryst.* A**32**, 832–847.

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. & Rossmann, M. G. (1974). *Structure determination of crystalline lobster* D*-glyceraldehyde-3-phosphate dehydrogenase. J. Mol. Biol.* **82**, 563–585.

Chothia, C. (1973). *Conformation of twisted β-pleated sheets in proteins. J. Mol. Biol.* **75**, 295–302.

Cochran, W., Crick, F. H. C. & Vand, V. (1952). *The structure of synthetic polypeptides. I. The transform of atoms on a helix. Acta Cryst.* **5**, 581–586.

Crowfoot, D. (1948). *X-ray crystallographic studies of compounds of biochemical interest. Annu. Rev. Biochem.* **17**, 115–146.

Crowfoot, D., Bunn, C. W., Rogers-Low, B. W. & Turner-Jones, A. (1949). *The X-ray crystallographic investigation of the structure of penicillin.* In *The Chemistry of Penicillin*, edited by H. T. Clarke, J. R. Johnson & R. Robinson, pp. 310–367. Princeton University Press.

Crowther, R. A. (1969). *The use of non-crystallographic symmetry for phase determination. Acta Cryst.* B**25**, 2571–2580.

Crowther, R. A. (1972). *The fast rotation function.* In *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1962). *The structure of haemoglobin. IX. A three-dimensional Fourier synthesis at 5.5 Å resolution: description of the structure. Proc. R. Soc. London Ser. A*, **265**, 161–187.

Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. (1985). *Structure of the protein subunits in the photosynthetic reaction centre of Rhodopseudomonas viridis at 3 Å resolution. Nature (London)*, **318**, 618–624.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *The crystal structure of myoglobin: phase determination to a resolution of 2 Å by the method of isomorphous replacement. Acta Cryst.* **14**, 1188–1195.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A. & Margoliash, E. (1971). *Ferricytochrome c. I. General features of the horse and bonito proteins at 2.8 Å resolution. J. Biol. Chem.* **246**, 1511–1535.

Dodson, E., Harding, M. M., Hodgkin, D. C. & Rossmann, M. G. (1966). *The crystal structure of insulin. III. Evidence for a 2-fold axis in rhombohedral zinc insulin. J. Mol. Biol.* **16**, 227–241.

Drenth, J., Hol, W. G. J., Jansonius, J. N. & Koekoek, R. (1972). *A comparison of the three-dimensional structures of subtilisin BPN′ and subtilisin Novo. Cold Spring Harbor Symp. Quant. Biol.* **36**, 107–116.

Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M. & Wolthers, B. G. (1968). *Structure of papain. Nature (London)*, **218**, 929–932.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Hahn, Th. (2005). Editor. *International Tables for Crystallography*, Vol. A. *Space-Group Symmetry.* Heidelberg: Springer.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. Acta Cryst.* **9**, 1–9.

Harker, D. & Kasper, J. S. (1947). *Phases of Fourier coefficients directly from crystal diffraction data. J. Chem. Phys.* **15**, 882–884.

Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. (1978). *Tomato bushy stunt virus at 2.9 Å resolution. Nature (London)*, **276**, 368–373.

Hendrickson, W. A. (1991). *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science*, **254**, 51–58.

Hendrickson, W. A. & Teeter, M. M. (1981). *Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. Nature (London)*, **290**, 107–113.

Henry, N. F. M. & Lonsdale, K. (1952). Editors. *International Tables for X-ray Crystallography*, Vol. I. *Symmetry Groups.* Birmingham: Kynoch Press.

Hermann, C. (1935). Editor. *Internationale Tabellen zur Bestimmung von Kristallstrukturen*, 1. Band. *Gruppentheoretische Tafeln.* Berlin: Begrüder Borntraeger.

Hilton, H. (1903). *Mathematical Crystallography and the Theory of Groups of Movements.* Oxford: Clarendon Press.

Hodgkin, D. C., Kamper, J., Lindsey, J., MacKay, M., Pickworth, J., Robertson, J. H., Shoemaker, C. B., White, J. G., Prosen, R. J. & Trueblood, K. N. (1957). *The structure of vitamin B_{12}. I. An outline of the crystallographic investigation of vitamin B_{12}. Proc. R. Soc. London Ser. A*, **242**, 228–263.

Holm, L. & Sander, C. (1997). *New structure – novel fold? Structure*, **5**, 165–171.

Holmes, K. C., Stubbs, G. J., Mandelkow, E. & Gallwitz, U. (1975). *Structure of tobacco mosaic virus at 6.7 Å resolution. Nature (London)*, **254**, 192–196.

Hooke, R. (1665). *Micrographia*, p. 85. London: Royal Society.

Huygens, C. (1690). *Traité de la lumière.* Leiden. (English translation by S. P. Thompson, 1912. London: Macmillan and Co.)

Jones, T. A. (1978). *A graphics model building and refinement system for macromolecules. J. Appl. Cryst.* **11**, 268–272.

Judson, H. F. (1979). *The Eighth Day of Creation.* New York: Simon & Schuster.

Kartha, G., Bello, J. & Harker, D. (1967). *Tertiary structure of ribonuclease. Nature (London)*, **213**, 862–865.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature (London)*, **181**, 662–666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **185**, 422–427.

Kim, S. H., Quigley, G. J., Suddath, F. L., McPherson, A., Sneden, D., Kim, J. J., Weinzierl, J. & Rich, A. (1973). *Three-dimensional structure of yeast phenylalanine transfer RNA: folding of the polynucleotide chain. Science*, **179**, 285–288.

Kraut, J., Robertus, J. D., Birktoft, J. J., Alden, R. A., Wilcox, P. E. & Powers, J. C. (1972). *The aromatic substrate binding site in subtilisin BPN′ and its resemblance to chymotrypsin. Cold Spring Harbor Symp. Quant. Biol.* **36**, 117–123.

Linderström-Lang, K. U. & Schellman, J. A. (1959). *Protein structure and enzyme activity.* In *The Enzymes*, edited by P. D. Boyer, Vol. 1, 2nd ed., pp. 443–510. New York: Academic Press.

Lipson, H. & Cochran, W. (1953). *The Determination of Crystal Structures*, pp. 325–326. London: G. Bell and Sons Ltd.

Lonsdale, K. (1928). *The structure of the benzene ring. Nature (London)*, **122**, 810.

Lonsdale, K. (1936). *Structure Factor Tables.* London: Bell.

McLachlan, D. Jr & Champaygne, E. F. (1946). *A machine for the application of sand in making Fourier projections of crystal structures. J. Appl. Phys.* **17**, 1006–1014.

Main, P. & Rossmann, M. G. (1966). *Relationships among structure factors due to identical molecules in different crystallographic environments. Acta Cryst.* **21**, 67–72.

Matthews, B. W. (1966). *The determination of the position of anomalously scattering heavy atom groups in protein crystals. Acta Cryst.* **20**, 230–239.

Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Three-dimensional structure of tosyl-α-chymotrypsin. Nature (London)*, **214**, 652–656.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol.* **247**, 536–540.

North, A. C. T. (1965). *The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. Acta Cryst.* **18**, 212–216.

Olby, R. (1974). *The Path to the Double Helix.* Seattle: University of Washington Press.

Patterson, A. L. (1934). *A Fourier series method for the determination of the components of interatomic distances in crystals. Phys. Rev.* **46**, 372–376.

Patterson, A. L. (1935). *A direct method for the determination of the components of interatomic distances in crystals. Z. Kristallogr.* **90**, 517–542.

Pauling, L. & Corey, R. B. (1951). *Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl Acad. Sci. USA*, **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pepinsky, R. (1947). *An electronic computer for X-ray crystal structure analysis. J. Appl. Phys.* **18**, 601–604.

Perutz, M. F. (1946). *Trans. Faraday Soc.* **42**B, 187.

Perutz, M. F. (1949). *An X-ray study of horse methaemoglobin. II. Proc. R. Soc. London Ser. A*, **195**, 474–499.

Perutz, M. F. (1951*a*). *New X-ray evidence on the configuration of polypeptide chains. Nature (London)*, **167**, 1053–1054.

Perutz, M. F. (1951*b*). *The 1.5-A. reflextion from proteins and polypeptides. Nature (London)*, **168**, 653–654.

Perutz, M. F. (1954). *The structure of haemoglobin. III. Direct determination of the molecular transform. Proc. R. Soc. London Ser. A*, **225**, 264–286.

Perutz, M. F. (1956). *Isomorphous replacement and phase determination in non-centrosymmetric space groups. Acta Cryst.* **9**, 867–873.

Perutz, M. F. (1962). *Proteins and Nucleic Acids. Structure and Function.* Amsterdam: Elsevier Publishing Co.

Perutz, M. F. (1997). *Keilin and the molteno.* In *Selected Topics in the History of Biochemistry: Personal Reflections, V*, edited by G. Semenza & R. Jaenicke, pp. 57–67. Amsterdam: Elsevier.

Ponder, J. W. & Richards, F. M. (1987). *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol.* **193**, 775–791.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Conformation of polypeptides and proteins. Adv. Protein Chem.* **23**, 283–438.

Reeke, G. N., Hartsuck, J. A., Ludwig, M. L., Quiocho, F. A., Steitz, T. A. & Lipscomb, W. N. (1967). *The structure of carboxypeptidase A, VI. Some results at 2.0-Å resolution, and the complex with glycyl-tyrosine at 2.8-Å resolution. Proc. Natl Acad. Sci. USA*, **58**, 2220–2226.

Richards, F. M. (1968). *The matching of physical models to three-dimensional electron-density maps: a simple optical device. J. Mol. Biol.* **37**, 225–230.

Richardson, J. S. (1977). *β-Sheet topology and the relatedness of proteins. Nature (London)*, **268**, 495–500.

Robertson, J. M. (1935). *An X-ray study of the structure of phthalocyanines. Part I. The metal-free, nickel, copper, and platinum compounds. J. Chem. Soc.* pp. 615–621.

Robertson, J. M. (1936). *Numerical and mechanical methods in double Fourier synthesis. Philos. Mag.* **21**, 176–187.

Robertus, J. D., Ladner, J. E., Finch, J. T., Rhodes, D., Brown, R. S., Clark, B. F. C. & Klug, A. (1974). *Structure of yeast phenylalanine tRNA at 3 Å resolution. Nature (London)*, **250**, 546–551.

Rossmann, M. G. (1960). *The accurate determination of the position and shape of heavy-atom replacement groups in proteins. Acta Cryst.* **13**, 221–226.

Rossmann, M. G. (1961). *The position of anomalous scatterers in protein crystals. Acta Cryst.* **14**, 383–388.

Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Structure of a human common cold virus and functional relationship to other picornaviruses. Nature (London)*, **317**, 145–153.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Rossmann, M. G. & Blow, D. M. (1963). *Determination of phases by the conditions of non-crystallographic symmetry. Acta Cryst.* **16**, 39–45.

Rossmann, M. G., Moras, D. & Olsen, K. W. (1974). *Chemical and biological evolution of a nucleotide-binding protein. Nature (London)*, **250**, 194–199.

Sanger, F. & Thompson, E. O. P. (1953*a*). *The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochem. J.* **53**, 353–366.

Sanger, F. & Thompson, E. O. P. (1953*b*). *The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J.* **53**, 366–374.

Sanger, F. & Tuppy, H. (1951). *The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem. J.* **49**, 481–490.

Schoenflies, A. M. (1891). *Krystallsysteme und Krystallstruktur.* Leipzig: Druck und Verlag von B. G. Teubner.

Snow, C. P. (1934). *The Search.* Indianapolis and New York: Bobbs-Merrill Co.

Watson, J. D. (1954). *The structure of tobacco mosaic virus. I. X-ray evidence of a helical arrangement of sub-units around the longitudinal axis. Biochim. Biophys. Acta*, **13**, 10–19.

Watson, J. D. (1968). *The Double Helix.* New York: Atheneum.

Wyckoff, H. W., Hardman, K. D., Allewell, N. M., Inagami, T., Johnson, L. N. & Richards, F. M. (1967). *The structure of ribonuclease-S at 3.5 Å resolution. J. Biol. Chem.* **242**, 3984–3988.

references

International Tables for Crystallography (2012). Vol. F, Chapter 1.3, pp. 13–38.

Chapter 1.3. Macromolecular crystallography and medicine

W. G. J. Hol and C. L. M. J. Verlinde

1.3.1. Introduction

In the last hundred years, crystallography has contributed immensely to the expansion of our understanding of the atomic structure of matter as it extends into the three spatial dimensions in which we describe the world around us. At the beginning of this century, the first atomic arrangements in salts, minerals and low-molecular-weight organic and metallo-organic compounds were unravelled. Then, initially one by one, but presently as an avalanche, the molecules of life were revealed in full glory at the atomic level with often astonishing accuracy, beginning in the 1950s when fibre diffraction first helped to resolve the structure of DNA, later the structures of polysaccharides, fibrous proteins, muscle and filamentous viruses. Subsequently, single-crystal methods became predominant and structures solved in the 1960s included myoglobin, haemoglobin and lysozyme, all of which were heroic achievements by teams of scientists, often building their own X-ray instruments, pioneering computational methods, and improving protein purification and crystallization procedures. Quite soon thereafter, in 1978, the three-dimensional structures of the first viruses were determined at atomic resolution. Less than ten years later, the mechanisms and structures of membrane proteins started to be unravelled. Presently, somewhere between five and ten structures of proteins are solved each day, about 85% by crystallographic procedures and about 15% by NMR methods. It is quite possible that within a decade the Protein Data Bank (PDB; Bernstein *et al.*, 1977) will receive a new coordinate set for a protein, RNA or DNA crystal structure every half hour. The resolution of protein crystal structures is improving dramatically and the size of the structures tackled is sometimes enormous: a virus with over a thousand subunits has been solved at atomic resolution (Grimes *et al.*, 1995) and the structure of the ribosome is on its way (Ban *et al.*, 1999; Cate *et al.*, 1999; Clemons *et al.*, 1999).

Macromolecular crystallography, discussed here in terms of its impact on medicine, is clearly making immense strides owing to a synergism of progress in many scientific disciplines including:

(*a*) *Computer hardware and software*: providing unprecedented computer power as well as instant access to information anywhere on the planet *via* the internet.

(*b*) *Physics*: making synchrotron radiation available with a wide range of wavelengths, very narrow bandwidths and very high intensities.

(*c*) *Materials science and instrumentation*: revolutionizing X-ray intensity measurements, with currently available charge-coupled-device detectors allowing protein-data collection at synchrotrons in tens of minutes, and with pixel array detectors on the horizon which are expected to collect a complete data set from a typical protein within a few seconds.

(*d*) *Molecular biology*: allowing the cloning, overexpression and modification of genes, with almost miraculous ease in many cases, resulting in a wide variety of protein variants, thereby enabling crystallization of 'impossible' proteins.

(*e*) *Genome sequencing*: determining complete bacterial genomes in a matter of months. With several eukaryote genomes and the first animal genome already completed, and with the human genome expected to be completed to a considerable degree by 2000, protein crystallographers suddenly have an unprecedented choice of proteins to study, giving rise to the new field of structural genomics.

(*f*) *Biochemistry and biophysics*: providing a range of tools for rapid protein and nucleic acid purification by size, charge and affinity, and for characterization of samples by microsequencing, fluorescence, mass spectrometry, circular dichroism and dynamic light scattering procedures.

(*g*) *Chemistry, in particular combinatorial chemistry*: discovering by more and more sophisticated procedures high affinity inhibitors or binders to drug target proteins which are of great interest by themselves, while in addition such compounds tend to improve co-crystallization results quite significantly.

(*h*) *Crystallography itself*: constantly developing new tools including direct methods, multi-wavelength anomalous-dispersion phasing techniques, maximum-likelihood procedures in phase calculation and coordinate refinement, interactive graphics and automatic model-building programs, density-modification methods, and the extremely important cryo-cooling techniques for protein and nucleic acid crystals, to mention only some of the major achievements in the last decade.

Numerous aspects of these developments are treated in great detail in this volume of *International Tables.*

1.3.2. Crystallography and medicine

Knowledge of accurate atomic structures of small molecules, such as vitamin B_{12}, steroids, folates and many others, has assisted medicinal chemists in their endeavours to modify many of these molecules for the combat of disease. The early protein crystallographers were well aware of the potential medical implication of the proteins they studied. Examples are the studies of the oxygen-carrying haemoglobin, the messenger insulin, the defending antibodies and the bacterial-cell-wall-lysing lysozyme. Yet, even by the mid-1980s, there were very few crystallographic projects which had the explicit goal of arriving at pharmaceutically active compounds (Hol, 1986). Since then, however, we have witnessed an incredible increase in the number of projects in this area with essentially every major pharmaceutical company having a protein crystallography unit, while in academia and research institutions the potential usefulness of a protein structure is often combined with the novelty of the system under investigation. In one case, the HIV protease, it might well be that, worldwide, the structure has been solved over one thousand times – in complex with hundreds of different inhibitory compounds (Vondrasek *et al.*, 1997).

Impressive as these achievements are, this seems to be only the beginning of medicinal macromolecular crystallography. The completion of the human genome project will provide an irresistible impetus for '*human structural genomics*': the determination, as rapidly and systematically as possible, of as many human protein structures as possible. The genome sequences of most

Copyright © 2012 International Union of Crystallography

Table 1.3.3.1
Crystal structures and genetic diseases

Crystal structure	Disease	Reference
Acidic fibroblast growth factor receptor	Familial Pfeiffer syndrome	[1]
Alpha-1-antitrypsin	Alpha-1-antitrypsin deficiency	[2]
Antithrombin III	Hereditary thrombophilia	[3], [4]
Arylsulfatase A	Leukodystrophy	[5]
Aspartylglucosaminidase	Aspartylglusominuria	[6]
Beta-glucuronidase	Sly syndrome	[7]
Branched-chain alpha-keto acid dehydrogenase	Maple syrup urine syndrome, type Ia	[39]
Carbonic anhydrase II	Guibaud–Vainsel syndrome, Marble brain disease	[8]
p53	Cancer	[9], [10]
Ceruloplasmin	Hypoceruloplasminemia	[11]
Complement C3	C3 complement component 3 deficiency	[12]
Cystatin B	Progressive myoclonus epilepsy	[13]
Factor VII	Factor VII deficiency	[14]
Factor VIII	Factor VIII deficiency	[40]
Factor X	Factor X deficiency (Stuart–Prower factor deficiency)	[15]
Factor XIII	Factor XIII deficiency	[16]
Fructose-1,6-bisphosphate aldolase	Fructose intolerance (fructosemia)	[41]
Gelsolin	Amyloidosis V	[17]
Growth hormone	Growth hormone deficiency	[18]
Haemochromatosis protein HFE	Hereditary haemochromatosis	[19]
Haemoglobin	Beta-thalassemia, sickle-cell anaemia	[20]
Tyrosine hydroxylase	Hereditary Parkinsonism	[21]
Hypoxanthine–guanine phosphoribosyltransferase	Lesch–Nyhan syndrome	[22]
Insulin	Hyperproinsulinemia, diabetes	[42]
Isovaleryl–coenzyme A dehydrogenase	Isovaleric acid CoA dehydrogenase deficiency	[23]
Lysosomal protective protein	Galactosialidosis	[24]
Ornithine aminotransferase	Ornithine aminotransferase deficiency	[25]
Ornithine transcarbamoylase	Ornithine transcarbamoylase deficiency	[43]
p16INK4a tumour suppressor	Cancer	[26]
Phenylalanine hydroxylase	Phenylketonuria	[27]
Plasminogen	Plasminogen deficiency	[28], [29], [30]
Protein C	Protein C deficiency	[31]
Purine nucleotide phosphorylase	Purine nucleotide phosphorylase deficiency	[32]
Serum albumin	Dysalbuminemic hyperthyroxinemia	[33]
Superoxide dismutase (Cu, Zn-dependent)	Familial amyotrophical lateral sclerosis	[34]
Thrombin	Hypoprothrombinemia, dysprothrombinemia	[35]
Transthyretrin	Amyloidosis I	[36]
Triosephosphate isomerase	Triosephosphate isomerase deficiency	[37]
Trypsinogen	Hereditary pancreatitis	[38]

References: [1] Blaber *et al.* (1996); [2] Loebermann *et al.* (1984); [3] Carrell *et al.* (1994); [4] Schreuder *et al.* (1994); [5] Lukatela *et al.* (1998); [6] Oinonen *et al.* (1995); [7] Jain *et al.* (1996); [8] Liljas *et al.* (1972); [9] Cho *et al.* (1994); [10] Gorina & Pavletich (1996); [11] Zaitseva *et al.* (1996); [12] Nagar *et al.* (1998); [13] Stubbs *et al.* (1990); [14] Banner *et al.* (1996); [15] Padmanabhan *et al.* (1993); [16] Yee *et al.* (1994); [17] McLaughlin *et al.* (1993); [18] DeVos *et al.* (1992); [19] Lebron *et al.* (1998); [20] Harrington *et al.* (1997); [21] Goodwill *et al.* (1997); [22] Eads *et al.* (1994); [23] Tiffany *et al.* (1997); [24] Rudenko *et al.* (1995); [25] Shah *et al.* (1997); [26] Russo *et al.* (1998); [27] Erlandsen *et al.* (1997); [28] Mulichak *et al.* (1991); [29] Mathews *et al.* (1996); [30] Chang, Mochalkin *et al.* (1998); [31] Mather *et al.* (1996); [32] Ealick *et al.* (1990); [33] He & Carter (1992); [34] Parge *et al.* (1992); [35] Bode *et al.* (1989); [36] Blake *et al.* (1978); [37] Mande *et al.* (1994); [38] Gaboriaud *et al.* (1996); [39] Ævarsson *et al.* (2000); [40] Pratt *et al.* (1999); [41] Gamblin *et al.* (1990); [42] Bentley *et al.* (1976); [43] Shi *et al.* (1998).

major infectious agents will be completed five years hence, if not sooner. This is likely to be followed up by '*selected pathogen structural genomics*', which will provide a wealth of pathogen protein structures for the design of new pharmaceuticals and probably also for vaccines.

This overview, written in late 1999, aims to convey some feel of the current explosion of 'crystallography in medicine'. Ten, perhaps even five, years ago it might have been feasible to make an almost comprehensive list of all protein structures of potentially direct medical relevance. Today, this is virtually impossible. Here we mention only selected examples in the text with apologies to the crystallographers whose projects should also have been mentioned, and to the NMR spectroscopists and electron microscopists whose work falls outside the scope of this review. Tables 1.3.3.1 and 1.3.4.1 to 1.3.4.5 provide more information, yet do not claim to cover comprehensively this exploding field. Also, not all of the structures listed were determined with medical applications in mind, though they might be exploited for drug design one day. These tables show at the same time tremendous achievements as well as great gaps in our structural knowledge of proteins from humans and human pathogens.

1.3.3. Crystallography and genetic diseases

Presently, an immense number of genetic diseases have been characterized at the genetic level and archived in OMIM [On-line Mendelian Inheritance in Man. Center for Medical Genetics, Johns Hopkins University (Baltimore, MD) and the National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 1999. URL: http://www.ncbi.nlm.nih.gov/omim/], with many more discoveries to occur in the next decades. Biomolecular crystallography has been very successful in explaining the cause of numerous genetic diseases at the atomic level. The stories of sickle cell anaemia, thalassemias and other deficiencies of haemoglobin set the stage (Dickerson & Geis, 1983), followed by numerous other examples (Table 1.3.3.1). Given the frequent occurrence of mutations in humans, it is likely that for virtually every structure of a human protein, a number of genetic diseases can be rationalized at the atomic level. Two investigations from the authors' laboratory may serve as examples:

(i) The severity of various cases of galactosialidosis – a lysosomal storage disease – could be related to the predicted effects of the amino-acid substitutions on the stability of human protective protein cathepsin A (Rudenko *et al.*, 1998).

(ii) The modification of Tyr393α to Asn in the branched-chain 2-oxo acid dehydrogenase occurs at the interface of the α and β subunits in this $\alpha_2\beta_2$ heterotetramer, providing a nice explanation of the 'mennonite' variants of maple syrup urine disease (MSUD) (Ævarsson *et al.*, 2000).

Impressive as the insights obtained into the causes of diseases like these might be, there is almost a sense of tragedy associated with this detailed understanding of a serious, sometimes fatal, afflictions at the atomic and three-dimensional level: there is often so little one can *do* with this knowledge. There are at least two, very different, reasons for this. The first reason is that turning a malfunctioning protein or nucleic acid into one that functions properly is notoriously difficult. Treatment would generally require the oral use of small molecules that somehow counteract the effect of the mutation, *i.e. the administration of the small molecule has to result in a functional complex of the drug with the mutant protein*. This is in almost all cases far more difficult than finding compounds that *block* the activity of a protein or nucleic acid – which is the way in which most current drugs function. The second reason for the paucity of drugs for treating genetic diseases is very different in nature: the number of patients suffering from a particular mutation responsible for a genetic disease is very small in most cases. This means that market forces do not encourage funding the expensive steps of testing the toxicity and efficacy of potentially pharmaceutically active compounds. One of several exceptions is sickle cell anaemia, where significant efforts have been made to arrive at pharmaceutically active agents (Rolan *et al.*, 1993). In this case the mutation Glu6βVal leads to deoxyhaemoglobin polymerization *via* the hydrophobic valine. In spite of several ingenious approaches based on the allosteric properties of haemoglobin (Wireko & Abraham, 1991), no successful compound seems to be on the horizon yet for the treatment of sickle cell anaemia.

More recently, the spectacular molecular mechanisms underlying genetic serpin deficiency diseases have been elucidated. A typical example is α1-antitrypsin deficiency, which leads to cirrhosis and emphysema. Normal α1-antitrypsin, a serine protease inhibitor, exposes a peptide loop as a substrate for the cognate proteinase in its active but metastable conformation. After cleavage of the loop, the protease becomes trapped as an acyl-enzyme with the serpin, and the cleaved serpin loop inserts itself as the central strand of one of the serpin β-sheets, accompanied by a dramatic change in protein stability. In certain mutant serpins, however, the exposed loop is conformationally more metastable and occasionally inserts itself into the β-sheet of a neighbouring serpin molecule, thereby forming serpin polymers with disastrous consequences for the patient (Carrell & Gooptu, 1998). *In vitro*, the polymerization of α1-antitrypsin can be reversed with synthetic homologues of the exposed peptide loop (Skinner *et al.*, 1998). This approach might be useful for other 'conformational diseases', which include Alzheimer's and other neurodegenerative disorders.

Another frequently occurring genetic disease is cystic fibrosis. Here we face a more complex situation than that in the case of sickle cell anaemia: a range of different mutations causes a malfunctioning of the same ion channel, which, consequently, leads to a range of severity of the disease (Collins, 1992). Protein crystallography is currently helpful in an indirect way in alleviating the problems of cystic fibrosis patients, not by studying the affected ion channel itself, but by revealing the structure of leukocyte elastase (Bode *et al.*, 1986), an enzyme responsible for much of the cellular damage associated with cystic fibrosis (Birrer, 1995). On the basis of the elastase structure, inhibitors were developed to combat the effects of the impaired ion channel (Warner *et al.*, 1994). Also, structures of key enzymes of *Pseudomonas aeruginosa*, a bacterium affecting many cystic fibrosis patients, form a basis for the design of therapeutics to treat infections by this pathogen. Yet, to the best of our knowledge, no compound has been developed so far that repairs the malfunctioning ion channel.

However, in some cases there might be more opportunities than assumed so far. Several mutations leading to genetic diseases result in a lack of stability of the affected protein. In instances when the mutant protein is still stable enough to fold, small molecules could conceivably be discovered that bind 'anywhere' to a pocket of these proteins, thereby stabilizing the protein. The same small molecule could even be able to increase the stability of proteins with *different* mildly destabilizing mutations. Such an approach, though not trivial by any means, might be worth pursuing. Proof of principle of this concept has recently been provided for several unstable p53 mutants, where the *same* small molecule enhanced the stability of *different* mutants (Foster *et al.*, 1999).

Of course, mutations that destroy cofactor binding or active sites, or destroy proper recognition of partner proteins, will be extremely difficult to correct by small molecules targeting the affected protein. In such instances, gene therapy is likely to be the way by which our and the next generation may be able to improve the lives of future generations.

1.3.4. Crystallography and development of novel pharmaceuticals

The impact of detailed knowledge of protein and nucleic acid structures on the design of new drugs has already been significant, and promises to be of tremendous importance in the next decades. The first structure of a known major drug bound to a target protein was probably that of methotrexate bound to dihydrofolate reductase (DHFR) (Matthews *et al.*, 1977). Even though the source of the enzyme was bacterial while methotrexate is used as a human anticancer agent, this protein–drug complex structure was nevertheless a hallmark achievement. It is generally accepted that the first protein-structure-inspired drug actually reaching the market was captopril, which is an antihypertensive compound blocking the action of angiotensin-converting enzyme, a metalloprotease. In this case, the structure of zinc-containing carboxypeptidase A was a guide to certain aspects of the chemical modification of lead compounds (Cushman & Ondetti, 1991). This success has been followed up by numerous projects specifically aimed at the design of new inhibitors, or activators, of carefully selected drug targets.

Structure-based drug design (SBDD) (Fig. 1.3.4.1) is the subject of several books and reviews that summarize projects and several success stories up until the mid-1990s (Kuntz, 1992; Perutz, 1992; Verlinde & Hol, 1994; Whittle & Blundell, 1994; Charifson, 1997; Veerapandian, 1997). Possibly the most dramatic impact made by SBDD has been on the treatment of AIDS, where the development of essentially all of the protease inhibitors on the market in 1999 has been guided by, or at least assisted by, the availability of numerous crystal structures of protease–inhibitor complexes.

The need for a large number of structures is common in all drug design projects and is due to several factors. One is the tremendous challenge for theoretical predictions of the correct binding mode and affinity of inhibitors to proteins. The current force fields are approximate, the properties of water are treacherous, the flexibility of protein and ligands lead quickly to a combinatorial explosion, and the free-energy differences between various binding modes are small. All this leads to the need for several experimental structures in a structure-based drug design cycle (Fig. 1.3.4.1). In this cycle, numerous disciplines are interacting in multiple ways. Many institutions, small and large, are following in one way or another this paradigm to speed up the lead discovery, lead optimization and even the bioavailability improvement steps in the drug development process. Moreover, a very powerful synergism exists between combinatorial chemistry and structure-based drug design. Structure-guided combinatorial libraries can utilize knowledge of ligand target sites in the design of the library [see *e.g.* Ferrer *et al.* (1999), Eckert *et al.* (1999) and Minke *et al.* (1999)]. Once tight-binding ligands are found by combinatorial methods, crystal structures of library compound–target complexes provide detailed information for new highly specific libraries.

The fate of a drug candidate during clinical tests can hinge on a single methyl group – just as a point mutation can alter the benefit of a wild-type protein molecule into the nightmare of a life-long genetic disease. Hence, many promising inhibitors eventually fail to be of benefit to patients. Nevertheless, knowledge of a *series* of protein structures in complex with inhibitors is of immense value in the design and development of future pharmaceuticals. In the following sections some examples will be looked at.

1.3.4.1. *Infectious diseases*

1.3.4.1.1. *Viral diseases*

Some icosahedral pathogenic viruses have all their capsid proteins elucidated, while for the more complex viruses like influenza virus, hepatitis C virus (HCV) and HIV, numerous individual protein structures have been solved (Table 1.3.4.1). However, not all 14 native proteins of the HIV genome have yet surrendered to the crystallographic community, nor to the NMR spectroscopists or the high-resolution electron microscopists, our partners in experimental structural biology (Turner & Summers, 1999). Nevertheless, the structures of HIV protease, reverse transcriptase and fragments of HIV integrase and of HIV viral core and surface proteins are of tremendous value for developing novel anti-AIDS therapeutics [Arnold *et al.*, 1996; Lin *et al.*, 1998; Wlodawer & Vondrasek, 1998; see also references in Table 1.3.4.1(*a*)]. A similar situation occurs for hepatitis C virus. The protease structure of this virus has been solved recently (simultaneously by four groups!), as well as its helicase structure, providing platforms on the basis of which the design of novel drugs is actively pursued (Le *et al.*, 1998).

A quite spectacular example of how structural knowledge can lead to the synthesis of powerful inhibitors is provided by influenza virus neuraminidase. The structure of a neuraminidase–transition-state analogue complex suggested the addition of positively charged amino and guanidinium groups to the C4 position of the analogue, which resulted, in one step, in a gain of four orders of magnitude in binding affinity for the target enzyme (von Itzstein *et al.*, 1993).

1.3.4.1.2. *Bacterial diseases*

A very large number of structures of important drug target proteins of bacterial origin have been solved crystallographically (Table 1.3.4.2). Currently, the most important single infectious bacterial agent is *Mycobacterium tuberculosis*, with three million deaths and eight million new cases annually (Murray & Salomon, 1998). The crystal structures of several *M. tuberculosis* potential and proven drug target proteins have been elucidated (Table 1.3.4.2). The complete *M. tuberculosis* genome has been sequenced recently (Cole *et al.*, 1998), and this will undoubtedly have a tremendous impact on future drug development.

The crystal structures of many bacterial dihydrofolate reductases, the target of several antimicrobials including trimethoprim, have also been reported. Recently, the atomic structure of dihydropteroate synthase (DHPS), the target of sulfa drugs, has been elucidated, almost 60 years after the first sulfa drugs were used to treat patients (Achari *et al.*, 1997; Hampele *et al.*, 1997).

A very special set of bacterial proteins are the toxins. Some of these have dramatic effects, with even a single molecule able to kill a host cell. These toxins outsmart and (mis)use many of the defence systems of the host, and their structures are often most unusual and fascinating, as recently reviewed by Lacy & Stevens (1998). The structures of the toxins are actively used for the design of prophylactics and therapeutic agents to treat bacterial diseases [see *e.g.* Merritt *et al.* (1997), Kitov *et al.* (2000) and Fan *et al.* (2000)]. It is remarkable that the properties of these devastating toxins are also used, or at least considered, for the treatment of disease, such as in the engineering of diphtheria toxin to create immunotoxins for the treatment of cancer and the deployment of cholera toxin mutants as adjuvants in mucosal vaccines. Knowledge of the three-dimensional structures of these toxins assists in the design of new therapeutically useful proteins.

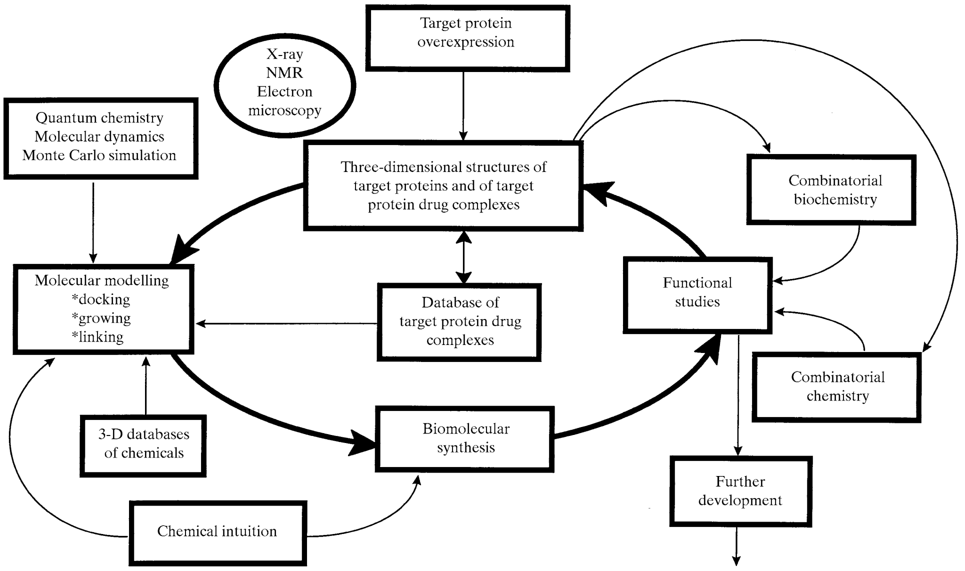


Figure 1.3.4.1
The structure-based drug design cycle.

1.3.4.1.3. *Protozoan infections*

A major cause of death and worldwide suffering is due to infections by several protozoa, including:

Table 1.3.4.1
Important human pathogenic viruses and their proteins

(*a*) RNA viruses
(i) Single-stranded

Family	Example	Protein structures solved	Reference
Arenaviridae	Lassa fever virus	None	
Bunyaviridae	Hantavirus	None	
Caliciviridae	Hepatitis E virus, Norwalk virus	None	
Coronaviridae	Corona virus	None	
Deltaviridae	Hepatitis D virus	Oligomerization domain of antigen	[1]
Filoviridae	Ebola virus	GP2 of membrane fusion glycoprotein	[2]
Flaviviridae	Dengue	NS3 protease	[3]
	Hepatitis C	NS3 protease	[4], [5]
		RNA helicase	[6]
	Yellow fever	None	
	Tick-borne encephalitis virus	Envelope glycoprotein	[7]
Orthomyxoviridae	Influenza virus	Neuraminidase	[8]
		Haemagglutinin	[9]
		Matrix protein M1	[10]
Paramyxoviridae	Measles, mumps, parainfluenza, respiratory syncytial virus	None	
Picornaviridae	Hepatitis A virus	3C protease	[11]
	Poliovirus	Capsid	[12]
		RNA-dependent polymerase	[13]
	Rhinovirus	Capsid	[14]
		3C protease	[15]
	Echovirus	Capsid	[16]
Retroviridae	HIV	Capsid protein	[17]
		Matrix protein	[18]
		Protease	[19], [20], [21]
		Reverse transcriptase	[22], [23], [47], [48], [49]
		Integrase	[24]
		gp120	[25]
		NEF	[26]
		gp41	[27]
Rhabdovirus	Rabies virus	None	
Togaviridae	Rubella	None	

(ii) Double-stranded

Family	Example	Protein structures solved	Reference
Reoviridae	Rotavirus	None	

(*b*) DNA viruses
(i) Single-stranded

Family	Example	Protein structures solved	Reference
Parvoviridae	B 19 virus	None	

(*a*) *Plasmodium falciparum* and related species, causing various forms of malaria;
(*b*) *Trypanosoma cruzi*, the causative agent of Chagas' disease in Latin America;
(*c*) *Trypanosoma brucei*, causing sleeping sickness in Africa;
(*d*) Some eleven different *Leishmania* species, responsible for several of the most horribly disfiguring diseases known to mankind.

Drug resistance, combined with other factors, has been the cause of a major disappointment for the early hopes of a 'malaria eradication campaign'. Fortunately, new initiatives have been launched recently under the umbrella of the 'Malaria roll back' program and the 'Multilateral Initiative for Malaria' (MIM). We are facing a formidable challenge, however, since the parasite is very clever at evading the immune response of the human host. Drugs are the mainstay of current treatments and may well be so for a long time to come. Protein crystallographic studies of *Plasmodium* proteins are hampered by the unusual codon usage of the *Plasmodium* species, coupled with a tendency to contain insertions of numerous hydrophilic residues in the polypeptide chain (Gardner *et al.*, 1998) which provide sometimes serious obstacles to obtaining large amounts of *Plasmodium* proteins for structural investigations.

However, the structures of an increasing number of potential drug targets from these protozoan parasites are being unravelled. These include the variable surface glycoproteins (VSGs) and glycolytic enzymes of *Trypanosoma brucei*, crucial malaria proteases, and the remarkable trypanothione reductase (Table 1.3.4.3). The structures of nucleotide phosphoribosyl transferases of a variety of protozoan parasites have also been elucidated. Moreover, the structure of DHFR from *Pneumocystis carinii*, the major opportunistic pathogen in AIDS patients in the United States, has been determined. Several of these structures are serving as starting points for the development of new drugs.

1.3.4.1.4. Fungi

In general, the human immune system is able to keep the growth of fungi under control, but in immuno-compromised

Table 1.3.4.1 (continued)

(ii) Double-stranded

Family	Example	Protein structures solved	Reference
Adenoviridae	Adenovirus	Protease	[28]
		Capsid	[29]
		Knob domain of fibre protein	[30]
Hepadnaviridae	Hepatitis B	Capsid	[31]
Herpesviridae	Cytomegalovirus	Protease	[32], [33], [34]
	Epstein–Barr virus	Domains of nuclear antigen 1	[35]
		BCRF1	[36]
	Herpes simplex	Protease	[37]
		Thymidine kinase	[38]
		Uracyl-DNA glycosylase	[39]
		Core of VP16	[40]
	Varicella zoster	Protease	[42]
Papovaviridae	Papillomavirus	DNA-binding domain of E2	[43]
		Activation domain of E2	[44]
Poxviridae	Smallpox virus	None	
	Vaccinia virus (related to smallpox but non-pathogenic)	Methyltransferase VP39	[45]
		Domain of topoisomerase	[46]

References: [1] Zuccola *et al.* (1998); [2] Weissenhorn *et al.* (1998); [3] Murthy *et al.* (1999); [4] Love *et al.* (1996); [5] Yan *et al.* (1998); [6] Yao *et al.* (1997); [7] Rey *et al.* (1995); [8] Varghese *et al.* (1983); [9] Wilson *et al.* (1981); [10] Sha & Luo (1997); [11] Allaire *et al.* (1994); [12] Hogle *et al.* (1985); [13] Hansen *et al.* (1997); [14] Rossmann *et al.* (1985); [15] Matthews *et al.* (1994); [16] Filman *et al.* (1998); [17] Worthylake *et al.* (1999); [18] Hill *et al.* (1996); [19] Navia, Fitzgerald *et al.* (1989); [20] Wlodawer *et al.* (1989); [21] Erickson *et al.* (1990); [22] Rodgers *et al.* (1995); [23] Ding *et al.* (1995); [24] Dyda *et al.* (1994); [25] Kwong *et al.* (1998); [26] Lee *et al.* (1996); [27] Chan *et al.* (1997); [28] Ding *et al.* (1996); [29] Roberts *et al.* (1986); [30] Xia *et al.* (1994); [31] Wynne *et al.* (1999); [32] Tong *et al.* (1996); [33] Qiu *et al.* (1996); [34] Shieh *et al.* (1996); [35] Bochkarev *et al.* (1995); [36] Zdanov *et al.* (1997); [37] Hoog *et al.* (1997); [38] Wild *et al.* (1995); [39] Savva *et al.* (1995); [40] Liu *et al.* (1999); [42] Qiu *et al.* (1997); [43] Hegde & Androphy (1998); [44] Harris & Botchan (1999); [45] Hodel *et al.* (1996); [46] Sharma *et al.* (1994); [47] Kohlstaedt *et al.* (1992); [48] Jacobo-Molina *et al.* (1993); [49] Ren *et al.* (1995).

patients (*e.g.* as a result of cancer chemotherapy, HIV infection, transplantation patients receiving immunosuppressive drugs, genetic disorders) such organisms are given opportunities they usually do not have. *Candida albicans* is an opportunistic fungal organism which causes serious complications in immunocompromised patients. Several of its proteins have been structurally characterized (Table 1.3.4.3) and provide opportunities for the development of selectively active inhibitors.

1.3.4.1.5. *Helminths*

Worms affect the lives of billions of human beings, causing serious morbidity in many ways. *Onchocerca volvolus* is the cause of river blindness, which resulted in the virtual disappearance of entire villages in West Africa, until ivermectin appeared. This remarkable compound dramatically reduced the incidence of the disease, even though it does not kill the adult worms. Therefore, a biological clock is ticking, waiting until resistance occurs against this single compound available for treatment. *Schistosoma* species are responsible for a wide variety of liver diseases and are spreading continuously since irrigation schemes provide a perfect environment for the intermediate snail vector. Other medically important helminths are *Wuchereria bancrofti* and *Brugia malayi.* Only a few protein structures from these very important disease-causing organisms have been unravelled so far (Table 1.3.4.3).

1.3.4.2. *Resistance*

Resistance to drugs in infectious organisms, as well as in cancers, is a fascinating subject, since it demonstrates the action and reaction of biological systems in response to environmental challenges. Life, of course, has been evolving to do just that – and the arrival of new chemicals, termed ‘drugs’, on the scene is nothing new to organisms that are the result of evolutionary processes involving billions of years of chemical warfare. Populations of organisms span a wide range of variation at the genetic and protein levels, and the chance that one of the variants is able to cope with drug pressure is nonzero. The variety of mechanisms observed to be responsible for drug resistance is impressive (Table 1.3.4.4).

Crystallography has made major contributions to the detailed molecular understanding of resistance in the case of detoxification, mutation and enzyme replacement mechanisms. Splendid examples are:

(*a*) The beta-lactamases: These beta-lactam degrading enzymes, of which there are four classes, are produced by many bacteria to counteract penicillins and cephalosporins, the most widely used antibiotics on the planet. No less than 53 structures of these enzymes reside in the PDB.

(*b*) HIV protease mutations: Tens of mutations have been characterized structurally. Many alter the active site at the site of mutation, thereby preventing drug binding. Other mutations rearrange the protein backbone, reshaping entire pockets in the binding site (Erickson & Burt, 1996).

(*c*) HIV reverse transcriptase mutations: *Via* structural studies, at least three mechanisms of drug resistance have been elucidated: direct alteration of the binding sites for the nucleoside analogue or non-nucleoside inhibitors, mutations that change the position of the DNA template, and mutations that induce conformational changes that propagate into the active site (Das *et al.*, 1996; Hsiou *et al.*, 1998; Huang, Chopra *et al.*, 1998; Ren *et al.*, 1998; Sarafianos *et al.*, 1999).

(*d*) Resistance to vancomycin: In non-resistant bacteria, vancomycin stalls the cell-wall synthesis by binding to the D-Ala-D-Ala terminus of the lipid–PP-disaccharide–pentapeptide substrate of the bacterial transglycosylase/transpeptidase, thereby sterically preventing the approach of the substrate. Resistant bacteria, however, have acquired a plasmid-borne transposon encoding for five genes, *vanS*, *vanR*, *vanH*, *vanA* and *vanX*, that allows them to synthesise a substrate ending in D-Ala-D-lactate. This minute difference, an oxygen atom replacing an NH, leads to a 1000-fold reduced affinity for vancomycin, explaining the resistance (Walsh *et al.*, 1996). Thus far, the structures of vanX (Bussiere *et al.*, 1998) and D-Ala-D-Ala ligase as a model for vanA (Fan *et al.*, 1994) have

Table 1.3.4.2
Protein structures of important human pathogenic bacteria

Organism	Disease(s)	Protein structures solved	Reference
Staphylococcus aureus	Abscesses	Alpha-haemolysin	[1]
	Endocarditis	Aureolysin	[2]
	Gastroenteritis	Beta-lactamase	[3]
	Toxic shock syndrome	Collagen adhesin	[4]
		7,8-Dihydroneopterin aldolase	[5]
		Dihydropteroate synthetase	[6]
		Enterotoxin A	[7]
		Enterotoxin B	[8]
		Enterotoxin C2	[9]
		Enterotoxin C3	[10]
		Exfoliative toxin A	[11]
		Ile-tRNA-synthetase	[12]
		Kanamycin nucleotidyltransferase	[13]
		Leukocidin F	[14]
		Nuclease	[15]
		Staphopain	[16]
		Staphylokinase	[17]
		Toxic shock syndrome toxin-1	[18]
Staphylococcus epidermidis	Implant infections	None	
Enterococcus faecalis	Urinary tract and biliary tract infections	NADH peroxidase	[19]
(*Streptococcus faecalis*)		Histidine-containing phosphocarrier protein	[20]
Streptococcus mutans	Endocarditis	Glyceraldehyde-3-phosphate dehydrogenase	[21]
Streptococcus pneumoniae	Pneumonia	Penicillin-binding protein PBP2x	[22]
	Meningitis, upper respiratory tract infections	Dpnm DNA adenine methyltransferase	[23]
Streptococcus pyogenes	Pharyngitis	Inosine monophosphate dehydrogenase	[24]
	Scarlet fever, toxic shock syndrome, immunologic disorders (acute glomerulonephritis and rheumatic fever)	Pyrogenic exotoxin C	[25]
Bacillus anthracis	Anthrax	Anthrax protective antigen	[26]
Bacillus cereus	Food poisoning	Beta-amylase	[27]
		Beta-lactamase II	[28]
		Neutral protease	[29]
		Oligo-1,6-glucosidase	[30]
		Phospholipase C	[31]
Clostridium botulinum	Botulism	Neurotoxin type A	[32]
Clostridium difficile	Pseudomembranous colitis	None	
Clostridium perfringens	Gas gangrene	Alpha toxin	[33]
	Food poisoning	Perfringolysin O	[34]
Clostridium tetani	Tetanus	Toxin C fragment	[35]
Corynebacterium diphtheriae	Diphtheria	Toxin	[36]
		Toxin repressor	[37]
Listeria monocytogenes	Meningitis, sepsis	Phosphatidylinositol-specific phospholipase C	[38]
Actinomyces israelii	Actinomycosis	None	
Nocardia asteroides	Nocardiosis	None	
Neisseria gonorrhoeae	Gonorrhea	Type 4 pilin	[39]
		Carbonic anhydrase	[40]
Neisseria meningitidis	Meningitis	Dihydrolipoamide dehydrogenase	[41]
Bordetella pertussis	Whooping cough	Toxin	[42]
		Virulence factor P.69	[43]
Brucella sp.	Brucellosis	None	
Campylobacter jejuni	Enterocolitis	None	
Enterobacter cloacae	Urinary tract infection, pneumonia	Beta-lactamase: class C	[44]
		UDP-*N*-acetylglucosamine enolpyruvyltransferase	[45]
Escherichia coli			
ETEC (enterotoxigenic)	Traveller's diarrhoea	Heat-labile enterotoxin	[46]
		Heat-stable enterotoxin (is a peptide)	[47]
EHEC (enterohaemorrhagic)	HUS	Verotoxin-1	[48]
EPEC (enteropathogenic)	Diarrhoea	None	
EAEC (enteroaggregative)	Diarrhoea	None	
EIEC (enteroinvasive)	Diarrhoea	None	
UPEC (uropathogenic)		FimH adhesin	[49]
		FimC chaperone	[49]
		PapD	[50]
NMEC (neonatal meningitis)	Meningitis	None	
Franciscella tularensis	Tularemia	None	
Haemophilus influenzae	Meningitis, otitis media, pneumonia	Diaminopimelate epimerase	[51]
		6-Hydroxymethyl-7,8-dihydropterin pyrophosphokinase	[52]
		Ferric iron binding protein Mirp	[53]
Klebsiella pneumoniae	Urinary tract infection, pneumonia, sepsis	β-Lactamase SHV-1	[54]
Legionella pneumophila	Pneumonia	None	
Pasteurella multocida	Wound infection	None	

Table 1.3.4.2 (continued)

Organism	Disease(s)	Protein structures solved	Reference
Proteus mirabilis	Pneumonia, urinary tract infection	Catalase	[55]
		Glutathione S-transferase	[56]
Proteus vulgaris	Urinary tract infections	Pvu II DNA-(cytosine N4) methyltransferase	[57]
		Pvu II endonuclease	[58]
		Tryptophanase	[59]
Salmonella typhi	Typhoid fever	None, but many for *S. typhimurium* linked with zoonotic disease	
Salmonella enteridis	Enterocolitis	None	
Serratia marcescens	Pneumonia, urinary tract infection	Serralysin	[60]
		Aminoglycoside 3-*N*-acetyltransferase	[61]
		Chitinase A	[62]
		Chitobiase	[63]
		Endonuclease	[64]
		Hasa (haemophore)	[65]
		Prolyl aminopeptidase	[66]
Shigella sp.	Dysentery	Chloramphenicol acetyltransferase	[67]
		Shiga-like toxin I	[68]
Vibrio cholerae	Cholera	Cholera toxin	[69], [70]
		DSBA oxidoreductase	[71]
		Neuraminidase	[104]
Yersinia enterocolitica	Enterocolitis	Protein-Tyr phosphatase YOPH	[72]
Yersinia pestis	Plague	None	
Pseudomonas aeruginosa	Wound infection, urinary tract infection, pneumonia, sepsis	Alkaline metalloprotease	[73]
		Amidase operon	[74]
		Azurin	[75]
		Cytochrome 551	[76]
		Cytochrome *c* peroxidase	[77]
		Exotoxin A	[78]
		p-Hydroxybenzoate hydroxylase	[79]
		Hexapeptide xenobiotic acetyltransferase	[80]
		Mandelate racemase	[81]
		Nitrite reductase	[82]
		Ornithine transcarbamoylase	[83]
		Porphobilinogen synthase	[84]
		Pseudolysin	[85]
Burkholderia cepacia	Wound infection, urinary tract infection, pneumonia, sepsis	Biphenyl-cleaving extradiol dioxygenase	[86]
		cis-Biphenyl-2,3-dihydrodiol-2,3-dehydrogenase	[87]
		Dialkylglycine decarboxylase	[88]
		Lipase	[89]
		Phthalate dioxygenase reductase	[90]
Stenotrophomonas maltophilia (= *Pseudomonas maltophilia*)	Sepsis	None	
Bacteroides fragilis	Intra-abdominal infections	Beta-lactamase type 2	[91]
Mycobacterium leprae	Leprosy	Chaperonin-10 (GroES homologue)	[92]
		RUVA protein	[93]
Mycobacterium tuberculosis	Tuberculosis	3-Dehydroquinate dehydratase	[94]
		Dihydrofolate reductase	[95]
		Dihydropteroate synthase	[96]
		Enoyl acyl-carrier-protein reductase (InhA)	[97]
		Mechanosensitive ion channel	[98]
		Quinolinate phosphoribosyltransferase	[99]
		Superoxide dismutase (iron dependent)	[100]
		Iron-dependent repressor	
Mycobacterium bovis	Tuberculosis	Tetrahydrodipicolinate-*N*-succinyltransferase	[102]
Chlamydia psitacci	Psittacosis	None	
Chlamydia pneumoniae	Atypical pneumonia	None	
Chlamydia trahomatis	Ocular, respiratory and genital infections	None	
Coxiella burnetii	Q fever	None	
Rickettsia sp.	Rocky Mountain spotted fever	None	
Borrelia burgdorferi	Lyme disease	Outer surface protein A	[103]
Leptospira interrogans	Leptospirosis	None	
Treponema pallidum	Syphilis	None	
Mycoplasma pneumoniae	Atypical pneumonia	None	

References: [1] Song *et al.* (1996); [2] Banbula *et al.* (1998); [3] Herzberg & Moult (1987); [4] Symersky *et al.* (1997); [5] Hennig *et al.* (1998); [6] Hampele *et al.* (1997); [7] Sundstrom *et al.* (1996); [8] Papageorgiou *et al.* (1998); [9] Papageorgiou *et al.* (1995); [10] Fields *et al.* (1996); [11] Vath *et al.* (1997); [12] Silvian *et al.* (1999); [13] Pedersen *et al.* (1995); [14] Pedelacq *et al.* (1999); [15] Loll & Lattman (1989); [16] Hofmann *et al.* (1993); [17] Rabijns *et al.* (1997); [18] Prasad *et al.* (1993); [19] Yeh *et al.* (1996); [20] Jia *et al.* (1993); [21] Cobessi *et al.* (1999); [22] Pares *et al.* (1996); [23] Tran *et al.* (1998); [24] Zhang, Evans *et al.* (1999); [25] Roussel *et al.* (1997); [26] Petosa *et al.* (1997); [27] Mikami *et al.* (1999); [28] Carfi *et al.* (1995); [29] Pauptit *et al.* (1988); [30] Watanabe *et al.* (1997); [31] Hough *et al.* (1989); [32] Lacy *et al.* (1998); [33] Naylor *et al.* (1998); [34] Rossjohn, Feil, McKinstry *et al.* (1997); [35] Umland *et al.* (1997); [36] Choe *et al.* (1992); [37] Qiu *et al.* (1995); [38] Moser *et al.* (1997); [39] Parge *et al.* (1995); [40] Huang, Xue *et al.* (1998); [41] Li de la Sierra *et al.* (1997); [42] Stein *et al.* (1994); [43] Emsley *et al.* (1996); [44] Lobkovsky *et al.* (1993); [45] Schonbrunn *et al.* (1996); [46] Sixma *et al.* (1991); [47] Ozaki *et al.* (1991); [48] Stein *et al.* (1992); [49] Choudhury *et al.* (1999); [50] Sauer *et al.* (1999); [51] Cirilli *et al.* (1993); [52] Hennig *et al.* (1999); [53] Bruns *et al.* (1997); [54] Kuzin *et al.* (1999); [55] Gouet *et al.* (1995); [56] Rossjohn, Polekhina *et al.* (1998); [57] Gong *et al.* (1997); [58] Athanasiadis *et al.* (1994); [59] Isupov *et al.* (1998); [60] Baumann (1994); [61] Wolf *et al.* (1998); [62] Perrakis *et al.* (1994); [63] Tews *et al.* (1996); [64] Miller *et al.* (1994); [65] Arnoux *et al.* (1999); [66] Yoshimoto *et al.* (1999); [67] Murray *et al.* (1995); [68] Ling *et al.* (1998); [69] Merritt *et al.* (1994); [70] Zhang *et al.* (1995); [71] Hu *et al.* (1997); [72] Stuckey *et al.* (1994); [73] Miyatake *et al.* (1995); [74] Pearl *et al.* (1994); [75] Adman *et al.* (1978); [76] Almassy & Dickerson (1978); [77] Fulop *et al.* (1995); [78] Allured *et al.* (1986); [79] Gatti *et al.* (1994); [80] Beaman *et al.* (1998); [81] Kallarakal *et al.* (1995); [82] Nurizzo *et al.* (1997); [83] Villeret *et al.* (1995); [84] Frankenberg *et al.* (1999); [85] Thayer *et al.* (1991); [86] Han *et al.* (1995); [87] Hulsmeyer *et al.* (1998); [88] Toney *et al.* (1993); [89] Kim *et al.* (1997); [90] Correll *et al.* (1992); [91] Concha *et al.* (1996); [92] Mande *et al.* (1996); [93] Roe *et al.* (1998); [94] Gourley *et al.* (1999); [95] Li *et al.* (2000); [96] Baca *et al.* (2000); [97] Dessen *et al.* (1995); [98] Chang, Spencer *et al.* (1998); [99] Sharma *et al.* (1998); [100] Cooper *et al.* (1995); [102] Beaman *et al.* (1997); [103] Li *et al.* (1997); [104] Crennell *et al*. (1994).

Table 1.3.4.3
Protein structures of important human pathogenic protozoa, fungi and helminths

(*a*) Protozoa

Organism	Disease	Protein structures solved	Reference
Acanthamoeba sp.	Opportunistic meningoencephalitis, corneal ulcers	Actophorin	[1]
		Profilin	[2]
Cryptosporidium parvum	Cryptosporidiosis	None	
Entamoeba histolytica	Amoebic dysentery, liver abscesses	None	
Giardia lamblia	Giardiasis	None	
Leishmania sp.	Leishmaniasis	Adenine phosphoribosyltransferase	[3]
		Dihydrofolate reductase-thymidylate synthase	[4]
		Glyceraldehyde-3-phosphate dehydrogenase	[5]
		Leishmanolysin	[6]
		Nucleoside hydrolase	[7]
		Pyruvate kinase	[8]
		Triosephosphate isomerase	[9]
Plasmodium sp.	Malaria	Fructose-1,6-bisphosphate aldolase	[10]
		Lactate dehydrogenase	[11]
		MSP1	[12]
		Plasmepsin II	[13]
		Purine phosphoribosyltransferase	[14]
		Triosephosphate isomerase	[15]
Pneumocystis carinii	Pneumonia	Dihydrofolate reductase	[16]
Toxoplasma gondii	Toxoplasmosis	HGXPRTase	[17]
		UPRTase	[18]
Trichomonas vaginalis	Trichomoniasis	None	
Trypanosoma brucei	Sleeping sickness	Fructose-1,6-bisphosphate aldolase	[19]
		Glyceraldehyde-3-phosphate dehydrogenase	[20]
		6-Phosphogluconate dehydrogenase	[21]
		Phosphoglycerate kinase	[22]
		Triosephosphate isomerase	[23]
		VSG	[24]
Trypanosoma cruzi	Chagas' disease	Cruzain (cruzipain)	[25]
		Glyceraldehyde-3-phosphate dehydrogenase	[26]
		Hypoxanthine phosphoribosyltransferase	[27]
		Triosephosphate isomerase	[28]
		Trypanothione reductase	[29]
		Tyrosine aminotransferase	[30]

(*b*) Fungi

Organism	Disease	Protein structures solved	Reference
Aspergillus fumigatus	Aspergillosis	Restrictocin	[31]
Blastomyces dermatidis	Blastomycosis	None	
Candida albicans	Candidiasis	Dihydrofolate reductase	[32]
		N-Myristoyl transferase	[33]
		Phosphomannose isomerase	[34]
		Secreted Asp protease	[35]
Coccidiodes immitis	Coccidioidomycosis	None	
Cryptococcus neoformans	Cryptococcosis	None	
Histoplasma capsulatum	Histoplasmosis	None	
Mucor sp.	Mucormycosis	None	
Paracoccidioides brasiliensis	Paracoccidioidomycosis	None	
Rhizopus sp.	Phycomycosis	Lipase II	[36]
		Rhizopuspepsin	[37]
		RNase Rh	[38]

been solved. They provide an exciting basis for arriving at new antibiotics against vancomycin-resistant bacteria.

(*e*) DHFR: Some bacteria resort to the 'ultimate mutation' in order to escape the detrimental effects of antibiotics. They simply replace the entire targeted enzyme by a functionally identical but structurally different enzyme. A prime example is the presence of a dimeric plasmid-encoded DHFR in certain trimethoprim-resistant bacteria. The structure proved to be unrelated to that of the chromosomally encoded monomeric DHFR (Narayana *et al.*, 1995).

Clearly, the structural insight gained from these studies provides us with avenues towards methods for coping with the rapid and alarming spread of resistance against available antibiotics that threatens the effective treatment of bacterial infections of essentially every person on this planet. This implies that we will constantly have to be aware of the potential occurrence of mono- and also *multi*-drug resistance, which has profound consequences for drug-design strategies for essentially all infectious diseases. It requires the development of *many* different compounds attacking many different target proteins and nucleic acids in the infectious agent. It appears to be crucial to use, from the very beginning, several new drugs *in combination* so that the chances of the occurrence of resistance are minimal. Multi-drug regimens have been spectacularly successful in the case of leprosy and HIV. Obviously, the development of vaccines is by far the

Table 1.3.4.3 (continued)

(*c*) Helminths

Organism	Disease	Protein structures solved	Reference
Clonorchis sinensis	Clonorchiasis	None	
Fasciola hepatica	Fasciolasis	Glutathione S-transferase	[39]
Fasciolopsis buski	Fasciolopsiasis	None	
Paragominus westermani	Paragonimiasis	None	
Schistosoma sp.	Schistosomiasis	Glutathione S-transferase Hexokinase	[39], [40] [41]
Diphyllobotrium latum	Diphyllobothriasis	None	
Echinococcus granulosus	Unilocular hydatid cyst disease	None	
Taenia saginata	Taeniasis	None	
Taenia solium	Taeniasis	None	
Ancylostoma duodenale	Old World hookworm disease	None	
Anisakis	Anisakiasis	None	
Ascaris lumbricoides	Ascariasis	Haemoglobin Major sperm protein Trypsin inhibitor	[42] [43] [44]
Enterobius vermicularis	Pinworm infection	None	
Necator	New World hookworm disease	None	
Strongyloides stercoralis	Strongyloidiasis	None	
Trichinella spiralis	Trichinosis	None	
Trichuris trichiura	Whipworm infection	None	
Brugia malayi	Filariasis	Peptidylprolyl isomerase	[45], [46]
Dracunculus medinensis	Guinea worm disease	None	
Loa loa	Loiasis	None	
Onchocerca volvulus	River blindness	None	
Toxocara canis	Visceral larva migrans	None	
Wuchereria bancrofti	Lymphatic filariasis (elephantiasis)	None	

References: [1] Leonard *et al.* (1997); [2] Liu *et al.* (1998); [3] Phillips *et al.* (1999); [4] Knighton *et al.* (1994); [5] Kim *et al.* (1995); [6] Schlagenhauf *et al.* (1998); [7] Shi, Schramm & Almo (1999); [8] Rigden *et al.* (1999); [9] Williams *et al.* (1999); [10] Kim *et al.* (1998); [11] Read *et al.* (1999); [12] Chitarra *et al.* (1999); [13] Silva *et al.* (1996); [14] Shi, Li *et al.* (1999); [15] Velanker *et al.* (1997); [16] Champness *et al.* (1994); [17] Schumacher *et al.* (1996); [18] Schumacher *et al.* (1998); [19] Chudzik *et al.* (2000); [20] Vellieux *et al.* (1993); [21] Phillips *et al.* (1998); [22] Bernstein *et al.* (1998); [23] Wierenga *et al.* (1987); [24] Freymann *et al.* (1990); [25] McGrath *et al.* (1995); [26] Souza *et al.* (1998); [27] Focia *et al.* (1998); [28] Maldonado *et al.* (1998); [29] Lantwin *et al.* (1994); [30] Blankenfeldt *et al.* (1999); [31] Yang & Moffat (1995); [32] Whitlow *et al.* (1997); [33] Weston *et al.* (1998); [34] Cleasby *et al.* (1996); [35] Cutfield *et al.* (1995); [36] Kohno *et al.* (1996); [37] Suguna *et al.* (1987); [38] Kurihara *et al.* (1992); [39] Rossjohn, Feil, Wilce *et al.* (1997); [40] McTigue *et al.* (1995); [41] Mulichak *et al.* (1998); [42] Yang *et al.* (1995); [43] Bullock *et al.* (1996); [44] Huang *et al.* (1994); [45] Mikol *et al.* (1998); [46] Taylor *et al.* (1998).

better solution, but it is not always possible. Antigenic variation, see *e.g.* the influenza virus, requires global vigilance and constant re-engineering of certain vaccines every year. Moreover, for higher organisms, and even for many bacterial species like *Shigella* (Levine & Noriega, 1995), with over 50 serotypes per species, the development of successful vaccines has, unfortunately, proved to be very difficult indeed. For sleeping sickness, the development of a vaccine is generally considered to be impossible. It is most likely, therefore, that world health will depend for centuries on a wealth of therapeutic drugs, together with many other measures, in order to keep the immense number of pathogenic organisms under control.

1.3.4.3. Non-communicable diseases

Of this large and diverse category of human afflictions we have already touched upon genetic disorders in Section 1.3.3 above. Other major types of non-communicable diseases include cancer, aging disorders, diabetes, arthritis, and cardiovascular and neurological illnesses. The field of non-communicable diseases is immense. Describing in any detail the current projects in, and potential impact of, protein and nucleic acid crystallography on these diseases would need more space than this entire volume on macromolecular crystallography. Hence, only a few selected examples out of the hundreds which could be described can be discussed here. Table 1.3.4.5 lists many examples of human protein structures elucidated without any claim as to completeness – it is simply impossible to keep up with the fountain of structures being determined at present. Yet, such tables do provide, it is hoped, an overview of what has been achieved and what needs to be done.

1.3.4.3.1. Cancers

Over a hundred different cancers have been described and clearly the underlying defect, loss of control of cell division, can be the result of many different shortcomings in different cells. The research in this area proceeds at a feverish pace, yet the development, discovery and design of effective but safe anti-cancer agents are unbelievably difficult challenges. The modifications needed to turn a normal cell into a malignant one are very small, hence the chance of arriving at 'true' anti-cancer drugs that exploit such small differences between normal and abnormal cells is exceedingly small. Nevertheless, such selective anti-cancer agents would leave normal cells essentially unaffected and are therefore the holy grail of cancer therapy. Few if any such compounds have been found so far, but cancer therapy is benefiting from a gradual increase in the number of useful compounds. Many have serious side effects, weaken the immune system and are barely tolerated by patients. However, they rescue large numbers of patients and hence it is of interest that many targets of these compounds, proteins and DNA molecules, have been structurally elucidated by crystallographic methods – often in complex with the cancer drug. The mode of action of many anti-cancer compounds is well understood, *e.g.* methotrexate targeting dihydrofolate reductase, and fluorouracil targeting thymidilate synthase. These are specific enzyme inhibitors acting along principles well known in other areas of medicine. Several anti-cancer drugs display unusual modes of action, such as:

(*a*) the DNA intercalators daunomycin (Wang *et al.*, 1987) and adriamycin (Zhang *et al.*, 1993);

(*b*) cisplatin, which forms DNA adducts (Giulian *et al.*, 1996);

Table 1.3.4.4
Mechanisms of resistance

Overexpress target protein
Mutate target protein
Use other protein with same function
Remove target altogether
Overexpress detoxification enzyme
Mutate detoxification enzyme
Create new detoxification enzyme
Mutate membrane porin protein
Remove or underexpress membrane porin protein
Overexpress efflux pumps
Mutate efflux pumps
Create/steal new efflux pumps
Improve DNA repair
Mutate prodrug conversion enzyme

(*c*) taxol, which not only binds to tubulin but also to bcl-2, thereby blocking the machinery of cancer cells in two entirely different ways (Amos & Lowe, 1999);

(*d*) camptothecin analogues, such as irinotecan and topotecan, which have the unusual property of prolonging the lifetime of a covalent topoisomerase–DNA complex, generating major road blocks on the DNA highway and causing DNA breakage and cell death;

(*e*) certain compounds function as minor-groove binders, *e.g.* netropsin and distamycin (Kopka *et al.*, 1985);

(*f*) completely new drugs which were developed based on the structures of matrix metalloproteinases, purine nucleotide phosphorylase and glycinamide ribonucleotide formyltransferase and which are in clinical trials (Jackson, 1997).

Meanwhile, it is sad that crystallography has not yet made any contribution to the molecular understanding of multi-drug resistance in cancer. The resistance is caused by cellular pumps that efficiently pump out the drugs, often leading to failed chemotherapy (Borst, 1999). On the other hand, the structures of major oncogenic proteins such as p21 (DeVos *et al.*, 1988; Pai *et al.*, 1989; Krengel *et al.*, 1990; Scheffzek *et al.*, 1997) and p53 (Cho *et al.*, 1994; Gorina & Pavletich, 1996) are of tremendous importance for future structure-based design of anti-neoplastic agents.

1.3.4.3.2. Diabetes

The hallmark characteristic of type I diabetes is a lack of insulin. A major therapeutic approach to this problem is insulin replacement therapy. Unfortunately, the insulin requirements of the body vary dramatically during the course of a day, with high concentrations needed at meal times and a basal level during the rest of the day. Only monomeric insulin is active at the insulin receptor level, but insulin has a natural tendency to form dimers and hexamers that dissociate upon dilution. Thanks to the three-dimensional insight obtained from dozens of insulin crystal structures, as wild-type (Hodgkin, 1971), mutants (Whittingham *et al.*, 1998) and in complex with zinc ions and small molecules such as phenol (Derewenda *et al.*, 1989), it has been possible to fine-tune the kinetics of insulin dissociation. The resulting availability of a variety of insulin preparations with rapid or prolonged action profiles has improved the quality of life of millions of people (Brange, 1997).

1.3.4.3.3. Blindness

The main causes of blindness worldwide are cataract, trachoma, glaucoma and onchocerciasis (Thylefors *et al.*, 1995). Trachoma and onchocerciasis are parasitic diseases that destroy the architecture of the eye; they were discussed in Section 1.3.4.1. The other two are discussed here. During cataract development, the lens of the eye becomes non-transparent as a result of aggregation of crystallins, preventing image formation. Crystal structures of several mammalian beta- and gamma-crystallins are known, but no human ones yet. In glaucoma, the optic nerve is destroyed by high intra-ocular pressure. One way to lower the pressure is to inhibit carbonic anhydrase II, a pivotal enzyme in maintaining the intra-ocular pressure. On the basis of the carbonic anhydrase crystal structure, researchers at Merck Research Laboratories were able to guide the optimization of an *S*-thienothiopyran-2-sulfonamide lead into a marketed drug for glaucoma: dorzolamide (Baldwin *et al.*, 1989).

1.3.4.3.4. Cardiovascular disorders

Thrombosis is a major cause of morbidity and mortality, especially in the industrial world. Hence, major effort is expended by pharmaceutical industries in the development of new classes of anti-coagulants with fewer side effects than available drugs, such as heparins and coumarins. Because blood coagulation is the result of an amplification cascade of enzymatic reactions, many potential targets are available. At present most of the effort is directed towards thrombin (Weber & Czarniecki, 1997) and factor Xa (Ripka, 1997), responsible for the penultimate step and the step immediately preceding it in the cascade, respectively. Thrombin is especially fascinating owing to the presence of at least three subsites: a primary specificity pocket with the catalytic serine-protease machinery, an exosite for recognizing extended fibrinogen and an additional pocket for binding heparin. This knowledge has led to the design of bivalent inhibitors which occupy two sites with ultra-high affinity and exquisite specificity. Several of these agents are in clinical trials (Pineo & Hull, 1999).

1.3.4.3.5. Neurological disorders

Even a quick glance at Table 1.3.4.5 shows that crystallography contributes to new therapeutics for numerous human afflictions and diseases. Yet there are major gaps in our understanding of protein functions, in particular of those involved in development and in neurological functions. These proteins are the target of many drugs obtained by classical pre-crystal-structure methods. These proven drug targets are very often membrane proteins involved in neuronal functions, and the diseases concerned are some of the most prevalent in mankind. A non-exhaustive list includes cerebrovascular disease (strokes), Parkinson's, epilepsy, schizophrenia, bipolar disease and depression.

Some of these diseases are heart-breaking afflictions, where parents have to accept the suicidal tendencies of their children, often with fatal outcomes; where partners have to endure the tremendous mood swings of their bipolar spouses and have to accept extreme excesses in behaviour; where a happy evening of life is turned into the gradual and sad demise of human intellect due to the progression of Alzheimer's, or to the loss of motor functions due to Parkinson's, or into the tragic stare of a victim of deep depression. Human nature, in all its shortcomings, has the tendency to try to help such tragic victims, but drugs for neurological disorders are rare, drug regimens are difficult to optimize and the commitment to follow a drug regimen – often for years, and often with major side effects – is a next to impossible task in many cases. New, better drugs are urgently needed and hence the structure determinations of the 'molecules of the brain' are major scientific as well as medical challenges of the next decades. Such

Table 1.3.4.5

Important human protein structures in drug design

Proteins from other species that might have been studied as substitutes for human ones were left out because of space limitations. We apologize to the researchers affected.

Pharmacological category	Protein	Reference
Synaptic and neuroeffector junctional function	None	
Central nervous system function	None	
Inflammation	Fibroblast collagenase (MMP-1) (also important in cancer)	[1], [2], [3], [4]
	Gelatinase	[5], [6]
	Stromelysin-1 (MMP-3) (also important in cancer)	[7], [8], [9], [10], [11]
	Matrilysin (MMP-7) (also important in cancer)	[12]
	Neutrophil collagenase (MMP-8) (also important in cancer)	[13], [14], [15], [16]
	Collagenase-3 (MMP-13)	[17]
	Human neutrophil elastase (also important for cystic fibrosis)	[18], [19], [20]
	Interleukin-1 beta converting enzyme (ICE)	[21], [22]
	p38 MAP kinase	[23], [24]
	Phospholipase A2	[25], [26], [27]
Renal and cardiovascular function	Renin	[28]
Gastrointestinal function	None	
Cancer	17-Beta-hydroxysteroid dehydrogenase	[29], [30]
	BRCT domain (BRCA1 C-terminus)	[31]
	Bcr-Abl kinase	[32]
	Cathepsin B	[33]
	Cathepsin D	[34], [35]
	CDK2	[36]
	CDK6	[37]
	DHFR	[38], [39]
	Acidic fibroblast growth factor (FGF)	[40]
	FGF receptor tyrosine kinase domain	[41]
	Glycinamide ribonucleotide formyl transferase	[42]
	Interferon-beta	[43]
	MMPs: see Inflammation	
	p53	[44], [45]
	p60 Src	[46]
	Purine nucleoside phosphorylase	[47]
	ras p21	[48], [49], [50], [51]
	Serine hydroxymethyltransferase	[52]
	S-Adenosylmethionine decarboxylase	[53]
	Thymidylate synthase	[54]
	Topoisomerase I	[55], [56]
	Tumour necrosis factor	[57]
	Interleukin 1-alpha	[58]
	Interleukin 1-beta	[59]
	Interleukin 1-beta receptor	[60], [61]
	Interleukin 8	[62]
Immunomodulation	Calcineurin	[63]
	Cathepsin S	[64]
	Cyclophilin	[65], [66], [67]
	Immunophilin FKBP12	[68], [69], [70]
	Inosine monophosphate dehydrogenase	[71]
	Interferon-gamma	[72], [73]
	Lymphocyte-specific kinase Lck	[74]
	PNP	[47]
	Syk kinase	[75]
	Tumour necrosis factor	[57]
	ZAP Tyr kinase	[76]
	Interleukin 2	[77]
	Interleukin 5	[78]
Haematopoiesis	Erythropoietin receptor	[79], [80]
Coagulation	AT III	[81], [82], [83], [84]
	Factor III	[85], [86]
	Factor VII	[87]
	Factor IX	[88]
	Factor X	[89]
	Factor XIII	[90]
	Factor XIV	[91]
	Fibrinogen: fragment	[92], [93]
	Plasminogen activator inhibitor (PAI)	[94], [95], [96]
	Thrombin	[97], [98], [99]
	tPA	[100]
	Urokinase-type plasminogen activator	[101]
	von Willebrand factor	[102], [103], [104]

Table 1.3.4.5 (continued)

Pharmacological category	Protein	Reference
Hormones and hormone receptors	Insulin	[105]
	Insulin receptor	[106], [107]
	Human growth hormone + receptor	[108]
	Oestrogen receptor	[109], [110]
	Progesterone receptor	[111]
	Prolactin receptor	[112]
Ocular function	Carbonic anhydrase	[113]
Genetic diseases	See Table 1.3.3.1	
Drug binding	Human serum albumin	[114], [115]
Drug metabolism	Glutathione S-transferase A-1	[116], [117]
	Glutathione S-transferase A4–4	[118]
	Glutathione S-transferase Mu-1	[119]
	Glutathione S-transferase Mu-2	[120]
Neurodegeneration	Aldose reductase	[121]
	JNK3	[122]
Osteoporosis	Cathepsin K	[123], [64]
	Src SH2	[126]
Various	Interferon-alpha 2b	[124]
	Bcl-xL	[125]

References: [1] Borkakoti *et al.* (1994); [2] Lovejoy, Cleasby *et al.* (1994); [3] Lovejoy, Hassell *et al.* (1994); [4] Spurlino *et al.* (1994); [5] Libson *et al.* (1995); [6] Gohlke *et al.* (1996); [7] Becker *et al.* (1995); [8] Dhanaraj *et al.* (1996); [9] Esser *et al.* (1997); [10] Gomis-Ruth *et al.* (1997); [11] Finzel *et al.* (1998); [12] Browner *et al.* (1995); [13] Bode *et al.* (1994); [14] Reinemer *et al.* (1994); [15] Stams *et al.* (1994); [16] Betz *et al.* (1997); [17] Gomis-Ruth *et al.* (1996); [18] Bode *et al.* (1986); [19] Wei *et al.* (1988); [20] Navia, McKeever *et al.* (1989); [21] Walker *et al.* (1994); [22] Rano *et al.* (1997); [23] Wilson *et al.* (1996); [24] Tong *et al.* (1997); [25] Scott *et al.* (1991); [26] Wery *et al.* (1991); [27] Kitadokoro *et al.* (1998); [28] Sielecki *et al.* (1989); [29] Ghosh *et al.* (1995); [30] Breton *et al.* (1996); [31] Zhang *et al.* (1998); [32] Nam *et al.* (1996); [33] Musil *et al.* (1991); [34] Baldwin *et al.* (1993); [35] Metcalf & Fusek (1993); [36] De Bondt *et al.* (1993); [37] Russo *et al.* (1998); [38] Oefner *et al.* (1988); [39] Davies *et al.* (1990); [40] Blaber *et al.* (1996); [41] McTigue *et al.* (1999); [42] Varney *et al.* (1997); [43] Karpusas *et al.* (1997); [44] Cho *et al.* (1994); [45] Gorina & Pavletich (1996); [46] Xu *et al.* (1997); [47] Ealick *et al.* (1990); [48] DeVos *et al.* (1988); [49] Pai *et al.* (1989); [50] Krengel *et al.* (1990); [51] Scheffzek *et al.* (1997); [52] Renwick *et al.* (1998); [53] Ekstrom *et al.* (1999); [54] Schiffer *et al.* (1995); [55] Redinbo *et al.* (1998); [56] Stewart *et al.* (1998); [57] Banner *et al.* (1993); [58] Graves *et al.* (1990); [59] Priestle *et al.* (1988); [60] Schreuder *et al.* (1997); [61] Vigers *et al.* (1997); [62] Baldwin *et al.* (1991); [63] Kissinger *et al.* (1995); [64] McGrath *et al.* (1998); [65] Kallen *et al.* (1991); [66] Ke *et al.* (1991); [67] Pfuegl *et al.* (1993); [68] Van Duyne, Standaert, Karplus *et al.* (1991); [69] Van Duyne, Standaert, Schreiber & Clardy (1991); [70] Van Duyne *et al.* (1993); [71] Colby *et al.* (1999); [72] Ealick *et al.* (1991); [73] Walter *et al.* (1995); [74] Zhu *et al.* (1999); [75] Futterer *et al.* (1998); [76] Meng *et al.* (1999); [77] Brandhuber *et al.* (1987); [78] Milburn *et al.* (1993); [79] Livnah *et al.* (1996); [80] Livnah *et al.* (1998); [81] Carrell *et al.* (1994); [82] Schreuder *et al.* (1994); [83] Skinner *et al.* (1997); [84] Skinner *et al.* (1998); [85] Muller *et al.* (1994); [86] Muller *et al.* (1996); [87] Banner *et al.* (1996); [88] Rao *et al.* (1995); [89] Padmanabhan *et al.* (1993); [90] Yee *et al.* (1994); [91] Mather *et al.* (1996); [92] Pratt *et al.* (1997); [93] Spraggon *et al.* (1997); [94] Mottonen *et al.* (1992); [95] Aertgeerts *et al.* (1995); [96] Xue *et al.* (1998); [97] Bode *et al.* (1989); [98] Rydel *et al.* (1990); [99] Rydel *et al.* (1994); [100] Laba *et al.* (1996); [101] Spraggon *et al.* (1995); [102] Bienkowska *et al.* (1997); [103] Huizinga *et al.* (1997); [104] Emsley *et al.* (1998); [105] Ciszak & Smith (1994); [106] Hubbard *et al.* (1994); [107] Hubbard (1997); [108] DeVos *et al.* (1992); [109] Schwabe *et al.* (1993); [110] Brzozowski *et al.* (1997); [111] Williams & Sigler (1998); [112] Somers *et al.* (1994); [113] Kannan *et al.* (1975); [114] He & Carter (1992); [115] Curry *et al.* (1998); [116] Sinning *et al.* (1993); [117] Cameron *et al.* (1995); [118] Bruns *et al.* (1999); [119] Tskovsky *et al.* (1999); [120] Raghunathan *et al.* (1994); [121] Wilson *et al.* (1992); [122] Xie *et al.* (1998); [123] Thompson *et al.* (1997); [124] Radhakrishnan *et al.* (1996); [125] Muchmore *et al.* (1996); [126] Waksman *et al.* (1993).

molecules will shed light on some of the deepest mysteries of humanity, including memory, cognition, desire, sleep *etc.* At the same time, such structures will provide opportunities for treating those suffering from neurodegenerative diseases due to age, genetic disposition, allergies, infections, traumas and combinations thereof. Such 'CNS protein structures' are one of the major challenges of biomacromolecular crystallography in the 21st century.

1.3.4.4. Drug metabolism and crystallography

As soon as a drug enters the body, an elaborate machinery comes into action to eliminate this foreign and potentially harmful molecule as quickly as possible. Two steps are usually distinguished in this process: phase I metabolism, in which the drug is functionalized, and phase II metabolism, in which further conjugation with endogenous hydrophilic molecules takes place, so that excretion *via* the kidneys can occur. Whereas this 'detoxification' process is essential for survival, it often renders promising inhibitors useless as drug candidates. Hence, structural knowledge of the proteins involved in metabolism could have a significant impact on the drug development process.

Thus far, only the structures of a few proteins crucial for drug distribution and metabolism have been elucidated. Human serum albumin binds hundreds of different drugs with micromolar dissociation constants, thereby altering drug levels in the blood dramatically. The structure of this important carrier molecule has been solved in complex with several drug molecules and should one day allow the prediction of the affinity of new chemical entities for this carrier protein, and thereby deepen our understanding of the serum concentrations of new candidate drugs (Carter & Ho, 1994; Curry *et al.*, 1998; Sugio *et al.*, 1999). Human oxidoreductases and hydrolases of importance in drug metabolism with known structure are: alcohol dehydrogenase (EC 1.1.1.1) (Hurley *et al.*, 1991), aldose reductase (EC 1.1.1.21) (Wilson *et al.*, 1992), glutathione reductase (NADPH) (EC 1.6.4.2) (Thieme *et al.*, 1981), catalase (EC 1.11.1.6) (Ko *et al.*, 2000), myeloperoxidase (EC 1.11.1.7) (Choi *et al.*, 1998) and beta-glucuronidase (EC 3.2.1.31) (Jain *et al.*, 1996). Recently, the first crystal structure of a mammalian cytochrome P-450, the most important class of xenobiotic metabolizing enzymes, has been reported (Williams *et al.*, 2000).

Of the conjugation enzymes, only glutathione S-transferases (EC 2.5.1.18) have been characterized structurally: A1 (Sinning *et al.*, 1993), A4-4 (Bruns *et al.*, 1999), MU-1 (Patskovsky *et al.*, 1999), MU-2 (Raghunathan *et al.*, 1994), P (Reinemer *et al.*, 1992) and THETA-2 (Rossjohn, McKinstry *et al.*, 1998). Tens of structures await elucidation in this area (Testa, 1994).

1.3.4.5. Drug manufacturing and crystallography

The development of drugs is a major undertaking and one of the hallmarks of modern societies. However, once a safe and effective therapeutic agent has been fully tested and approved, manufacturing the compound on a large scale is often the next major challenge. Truly massive quantities of penicillin and cephalosporin are produced worldwide, ranging from 2000 to 7000 tons annually (Conlon *et al.*, 1995). In the production of semi-synthetic penicillins, the enzyme penicillin acylase plays a very significant role. This enzyme catalyses the hydrolysis of penicillin into 6-aminopenicillanic acid. Its crystal structure has been elucidated (Duggleby *et al.*, 1995) and may now be used for protein-engineering studies to improve its properties for the biotechnology industry. The production of cephalosporins could benefit in a similar way from knowing the structure of cephalosporin acylase (CA), since the properties of this enzyme are not optimal for use in production plants. Therefore, the crystal structure determination of CA could provide a basis for improving the substrate specificity of CA by subsequent protein-engineering techniques. Fortunately, a first CA structure has been solved recently (Kim *et al.*, 2000), with many other structures expected to be solved essentially simultaneously. Clearly, crystallography can be not only a major player in the design and optimization of therapeutic drugs, but also in their manufacture.

1.3.5. Vaccines, immunology and crystallography

Vaccines are probably the most effective way of preventing disease. An impressive number of vaccines have been developed and many more are under development (National Institute of Allergy and Infectious Diseases, 1998). Smallpox has been eradicated thanks to a vaccine, and polio is being targeted for eradication in a worldwide effort, again using vaccination strategies. To the best of our knowledge, crystal structures of viruses, viral capsids or viral proteins have not been used in developing the currently available vaccines. However, there are projects underway that may change this.

For instance, the crystal structure of rhinovirus has resulted in the development of compounds that have potential as antiviral agents, since they stabilize the viral capsid and block, or at least delay, the uncoating step in viral cell entry (Fox *et al.*, 1986). These rhinovirus capsid-stabilizing compounds are, in a different project, being used to stabilize poliovirus particles against heat-induced denaturation in vaccines (Grant *et al.*, 1994). This approach may be applicable to other cases, although it has not yet resulted in commercially available vaccine-plus-stabilizer cocktails. However, it is fascinating to see how a drug-design project may be able to assist vaccine development in a rather unexpected manner.

Three-dimensional structural information about viruses is also being used to aid in the development of vaccines. Knowledge of the architecture of and biological functions of coat proteins has been used to select loops at viral surfaces that can be replaced with antigenic loops from other pathogens for vaccine-engineering purposes (*e.g.* Burke *et al.*, 1988; Kohara *et al.*, 1988; Martin *et al.*, 1988; Murray *et al.*, 1998; Arnold *et al.*, 1994; Resnick *et al.*, 1995; Smith *et al.*, 1998; Arnold & Arnold, 1999; Zhang, Geisler *et al.*, 1999). The design of human rhinovirus (HRV) and poliovirus chimeras has been aided by knowing the atomic structure of the viruses (Hogle *et al.*, 1985; Rossmann *et al.*, 1985; Arnold & Rossmann, 1988; Arnold & Rossmann, 1990) and detailed features of the neutralizing immunogenic sites on the virion surfaces (Sherry & Rueckert, 1985; Sherry *et al.*, 1986). In this way, one can imagine that in cases where the atomic structures of antigenic loops in 'donor' immunogens are known as well as the structure of the 'recipient' loop in the virus capsid protein, optimal loop transplantation might become possible. It is not yet known how to engineer precisely the desired three-dimensional structures and properties into macromolecules. However, libraries of macromolecules or viruses constructed using combinatorial mutagenesis can be searched to increase the likelihood of including structures with desired architecture and properties such as immunogenicity. With appropriate selection methods, the rare constructs with desired properties can be identified and 'fished out'. Research of this type has yielded some potently immunogenic presentations of sequences transplanted on the surface of HRV (reviewed in Arnold & Arnold, 1999). For reasons not quite fully understood, presenting multiple copies of antigens to the immune system leads to an enhanced immune response (Malik & Perham, 1997). It is conceivable that, eventually, it might even be possible for conformational epitopes consisting of multiple 'donor' loops to be grafted onto 'recipient capsids' while maintaining the integrity of the original structure. Certainly, such feats are difficult to achieve with present-day protein-engineering skills, but recent successes in protein design offer hope that this will be feasible in the not too distant future (Gordon *et al.*, 1999).

Immense efforts have been made by numerous crystallographers to unravel the structures of molecules involved in the unbelievably complex, powerful and fascinating immune system. Many of the human proteins studied are listed in Table 1.3.4.5 with, as specific highlights, the structures of immunoglobulins (Poljak *et al.*, 1973), major histocompatibility complex (MHC) molecules (Bjorkman *et al.*, 1987; Brown *et al.*, 1993; Fremont *et al.*, 1992; Bjorkman & Burmeister, 1994), T-cell receptors (TCR) and MHC:TCR complexes (Garboczi *et al.*, 1996; Garcia *et al.*, 1996), an array of cytokines and chemokines, and immune cell-specific kinases such as lck (Zhu *et al.*, 1999). This knowledge is being converted into practical applications, for instance by humanising non-human antibodies with desirable properties (Reichmann *et al.*, 1988) and by creating immunotoxins.

The interactions between chemokines and receptors, and the complicated signalling pathways within each immune cell, make it next to impossible to predict the effect of small compounds interfering with a specific protein–protein interaction in the immune system (Deller & Jones, 2000). However, great encouragement has been obtained from the discovery of the remarkable manner by which the immunosuppressor FK506 functions: this small molecule brings two proteins, FKB12 and calcineurin, together, thereby preventing T-cell activation by calcineurin. The structure of this remarkable ternary complex is known (Kissinger *et al.*, 1995). Such discoveries of unusual modes of action of therapeutic compounds are the foundation for new concepts such as 'chemical dimerizers' to activate signalling events in cells such as apoptosis (Clackson *et al.*, 1998).

In spite of the gargantuan task ahead aimed at unravelling the cell-to-cell communication in immune action, it is unavoidable that the next decades will bring us unprecedented insight into the many carefully controlled processes of the immune system. In turn, it is expected that this will lead to new therapeutics for manipulating a truly wonderful defence system in order to assist vaccines, to decrease graft rejection processes in organ transplants and to control auto-immune diseases that are likely to be playing a major role in cruelly debilitating diseases such as rheumatoid arthritis and type I diabetes.

1.3.6. Outlook and dreams

At the beginning of the 1990s, Max Perutz inspired many researchers with a passion for structure and a heart for the suffering of mankind with a fascinating book entitled *Protein Structure – New Approaches to Disease and Therapy* (Perutz, 1992). The explosion of medicinal macromolecular crystallography since then has been truly remarkable. What should we expect for the next decades?

In the realm of safe predictions we can expect the following:

(*a*) High-throughput macromolecular crystallography due to the developments outlined in Section 1.3.1, leading to the new field of 'structural genomics'.

(*b*) Crystallography of very large complexes. While it is now clear that an atomic structure of a complex of 58 proteins and three RNA molecules, the ribosome, is around the corner, crystallographers will widen their horizons and start dreaming of structures like the nuclear pore complex, which has a molecular weight of over 100 000 000 Da.

(*c*) A steady flow of membrane protein structures. Whereas Max Perutz could only list five structures in his book of 1992, there are now over 40 PDB entries for membrane proteins. Most of them are transmembrane proteins: bacteriorhodopsin, photoreaction centres, light-harvesting complexes, cytochrome bc_1 complexes, cytochrome *c* oxidases, photosystem I, porins, ion channels and bacterial toxins such as haemolysin and LukF. Others are monotopic membrane proteins such as squalene synthase and the cyclooxygenases. Clearly, membrane protein crystallography is gaining momentum at present and may open the door to atomic insight in neurotransmitter pharmacology in the next decade.

What if we dream beyond the obvious? One day, medicinal crystallography may contribute to:

(*a*) The design of submacromolecular agonists and antagonists of proteins and nucleic acids in a matter of a day by integrating rapid structure determinations, using only a few nanograms of protein, with the power of combinatorial and, in particular, computational chemistry.

(*b*) 'Structural toxicology' based on 'human structural genomics'. Once the hundreds of thousands of structures of human proteins and complexes with other proteins and nucleic acids have been determined, truly predictive toxicology may become possible. This will not only speed up the drug-development process, but may substantially reduce the suffering of animals in preclinical tests.

(*c*) The creation of completely new classes of drugs to treat addiction, organ regeneration, aging, memory enhancement *etc.*

One day, crystallography will have revealed the structure of hundreds of thousands of proteins and nucleic acids from human and pathogen, and their complexes with each other and with natural and designed low-molecular-weight ligands. This will form an extraordinarily precious database of knowledge for furthering the health of humans. Hence, in the course of the 21st century, crystallography is likely to become a major driving force for improving health care and disease prevention, and will find a well deserved place in future books describing progress in medicine, sometimes called 'The Greatest Benefit to Mankind' (Porter, 1999).

We wish to thank Heidi Singer for terrific support in preparing the manuscript, and Drs Alvin Kwiram, Michael Gelb, Seymour Klebanov, Wes Van Voorhis, Fred Buckner, Youngsoo Kim and Rein Zwierstra for valuable comments.

References

Achari, A., Somers, D. O., Champness, J. N., Bryant, P. K., Rosemond, J. & Stammers, D. K. (1997). *Crystal structure of the anti-bacterial sulfonamide drug target dihydropteroate synthase. Nature Struct. Biol.* **4**, 490–497.

Adman, E. T., Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1978). *A crystallographic model for azurin at 3 Å resolution. J. Mol. Biol.* **123**, 35–47.

Aertgeerts, K., De Bondt, H. L., De Ranter, C. J. & Declerck, P. J. (1995). *Mechanisms contributing to the conformational and functional flexibility of plasminogen activator inhibitor-1. Nature Struct. Biol.* **2**, 891–897.

Ævarsson, A., Chuang, J. L., Wynn, R. M., Turley, S., Chuang, D. T. & Hol, W. G. J. (2000). *Crystal structure of human branched-chain α-ketoacid dehydrogenase and the molecular basis of multienzyme complex deficiency in maple syrup urine disease. Struct. Fold. Des.* **8**, 277–291.

Allaire, M., Chernaia, M. M., Malcolm, B. A. & James, M. N. (1994). *Picornaviral 3C cysteine proteinases have a fold similar to chymotrypsin-like serine proteinases. Nature (London)*, **369**, 72–76.

Allured, V. S., Collier, R. J., Carroll, S. F. & McKay, D. B. (1986). *Structure of exotoxin A of Pseudomonas aeruginosa at 2.0-Å resolution. Proc. Natl Acad. Sci. USA*, **83**, 1320–1324.

Almassy, R. J. & Dickerson, R. E. (1978). *Pseudomonas cytochrome c551 at 2.0 Å resolution: enlargement of the cytochrome c family. Proc. Natl Acad. Sci. USA*, **75**, 2674–2678.

Amos, L. A. & Lowe, J. (1999). *How Taxol stabilises microtubule structure. Chem. Biol.* **6**, R65–R69.

Arnold, E., Das, K., Ding, J., Yadav, P. N., Hsiou, Y., Boyer, P. L. & Hughes, S. H. (1996). *Targeting HIV reverse transcriptase for anti-AIDS drug design: structural and biological considerations for chemotherapeutic strategies. Drug Des. Discov.* **13**, 29–47.

Arnold, E. & Rossmann, M. G. (1988). *The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. Acta Cryst.* A**44**, 270–283.

Arnold, E. & Rossmann, M. G. (1990). *Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. J. Mol. Biol.* **211**, 763–801.

Arnold, G. F. & Arnold, E. (1999). *Using combinatorial libraries to develop vaccines. ASM News*, **65**, 603–610.

Arnold, G. F., Resnick, D. A., Li, Y., Zhang, A., Smith, A. D., Geisler, S. C., Jacobo-Molina, A., Lee, W., Webster, R. G. & Arnold, E. (1994). *Design and construction of rhinovirus chimeras incorporating immunogens from polio, influenza, and human immunodeficiency viruses. Virology*, **198**, 703–708.

Arnoux, P., Haser, R., Izadi, N., Lecroisey, A., Delepierre, M., Wandersman, C. & Czjzek, M. (1999). *The crystal structure of HasA, a hemophore secreted by Serratia marcescens. Nature Struct. Biol.* **6**, 516–520.

Athanasiadis, A., Vlassi, M., Kotsifaki, D., Tucker, P. A., Wilson, K. S. & Kokkinidis, M. (1994). *Crystal structure of PvuII endonuclease reveals extensive structural homologies to EcoRV. Nature Struct. Biol.* **1**, 469–475.

Baca, A. M., Sirawaraporn, R., Turley, S., Athappilly, F., Sirawaraporn, W. & Hol, W. G. J. (2000). *Crystal structure of Mycobacterium tuberculosis 6-hydroxymethyl-1,8-dihydropteroate synthase in complex with pterin monophosphate: new insight into the enzymatic mechanism and sulfa-drug action. J. Mol. Biol.* **302**, 1193–1212.

Baldwin, E. T., Bhat, T. N., Gulnik, S., Hosur, M. V., Sowder, R. C. I., Cachau, R. E., Collins, J., Silva, A. M. & Erickson, J. W. (1993). *Crystal structures of native and inhibited forms of human cathepsin D: implications for lysosomal targeting and drug design. Proc. Natl Acad. Sci. USA*, **90**, 6796–6800.

Baldwin, E. T., Weber, I. T., St Charles, R., Xuan, J. C., Appella, E., Yamada, M., Matsushima, K., Edwards, B. F., Clore, G. M. & Gronenborn, A. M. (1991). *Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proc. Natl Acad. Sci. USA*, **88**, 502–506.

Baldwin, J. J., Ponticello, G. S., Anderson, P. S., Christy, M. E., Murcko, M. A., Randall, W. C., Schwam, H., Sugrue, M. F., Springer, J. P., Gautheron, P., Grove, J., Mallorga, P., Viader, M. P., McKeever, B. M. & Navia, M. A. (1989). *Thienothiopyran-2-sulfonamides: novel topically active carbonic anhydrase inhibitors for the treatment of glaucoma. J. Med. Chem.* **32**, 2510–2513.

Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. & Steitz, T. A. (1999). *Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. Nature (London)*, **400**, 841–847.

Banbula, A., Potempa, J., Travis, J., Fernandez-Catalan, C., Mann, K., Huber, R., Bode, W. & Medrano, F. (1998). *Amino-acid sequence and three-dimensional structure of the Staphylococcus aureus metalloproteinase at 1.72 Å resolution. Structure*, **6**, 1185–1193.

Banner, D. W., D'Arcy, A., Chene, C., Winkler, F. K., Guha, A., Konigsberg, W. H., Nemreson, Y. & Kirchhofer, D. (1996). *The crystal structure of the complex of blood coagulation factor VIIa with soluble tissue factor. Nature*, **380**, 41–46.

Banner, D. W., D'Arcy, A., Janes, W., Gentz, R., Schoenfeld, H. J., Broger, C., Loetscher, H. & Lesslauer, W. (1993). *Crystal structure of the soluble human 55 kd TNF receptor–human TNF beta complex: implications for TNF receptor activation. Cell*, **7**, 431–445.

Baumann, U. (1994). *Crystal structure of the 50 kDa metallo protease from Serratia marcescens. J. Mol. Biol.* **242**, 244–251.

Beaman, T. W., Binder, D. A., Blanchard, J. S. & Roderick, S. L. (1997). *Three-dimensional structure of tetrahydrodipicolinate N-succinyltransferase. Biochemistry*, **36**, 489–494.

Beaman, T. W., Sugantino, M. & Roderick, S. L. (1998). *Structure of the hexapeptide xenobiotic acetyltransferase from Pseudomonas aeruginosa. Biochemistry*, **37**, 6689–6696.

Becker, J. W., Marcy, A. I., Rokosz, L. L., Axel, M. G., Burbaum, J. J., Fitzgerald, P. M., Cameron, P. M., Esser, C. K., Hagmann, W. K., Hermes, J. D. & Springer, J. P. (1995). *Stromelysin-1: three dimensional structure of the inhibited catalytic domain and of the C-truncated proenzyme. Protein Sci.* **4**, 1966–1976.

Bentley, G., Dodson, E., Dodson, G., Hodgkin, D. & Mercola, D. (1976). *Structure of insulin in 4-zinc insulin. Nature (London)*, **261**, 166–168.

Bernstein, B. E., Williams, D. M., Bressi, J. C., Kuhn, P., Gelb, M. H., Blackburn, G. M. & Hol, W. G. J. (1998). *A bisubstrate analog induces unexpected conformational changes in phosphoglycerate kinase from Trypanosoma brucei. J. Mol. Biol.* **279**, 1137–1148.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur. J. Biochem.* **80**, 319–324.

Betz, M., Huxley, P., Davies, S. J., Mushtaq, Y., Pieper, M., Tschesche, H., Bode, W. & Gomis-Ruth, F. X. (1997). *1.8-Å crystal structure of the catalytic domain of human neutrophil collagenase (matrix metalloproteinase-8) complexed with a peptidomimetic hydroxamate primed-side inhibitor with a distinct selectivity profile. Eur. J. Biochem.* **247**, 356–363.

Bienkowska, J., Cruz, M., Atiemo, A., Handin, R. & Liddington, R. (1997). *The von Willebrand factor A3 domain does not contain a metal ion-dependent adhesion site motif. J. Biol. Chem.* **272**, 25162–25167.

Birrer, P. (1995). *Proteases and antiproteases in cystic fibrosis: pathogenetic considerations and therapeutic strategies. Respiration*, **62**, S25–S28.

Bjorkman, P. J. & Burmeister, W. P. (1994). *Structures of two classes of MHC molecules elucidated: crucial differences and similarities. Curr. Opin. Struct. Biol.* **4**, 852–856.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987). *Structure of human class I histocompatibility antigen, HLA-A2. Nature (London)*, **329**, 506–512.

Blaber, M., DiSalvo, J. & Thomas, K. A. (1996). *X-ray crystal structure of human acidic fibroblast growth factor. Biochemistry*, **35**, 2086–2094.

Blake, C. C., Geisow, M. J., Oatley, S. J., Rerat, B. & Rerat, C. (1978). *Structure of prealbumin: secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å. J. Mol. Biol.* **121**, 339–356.

Blankenfeldt, W., Nowicki, C., Montemartini-Kalisz, M., Kalisz, H. M. & Hecht, H. J. (1999). *Crystal structure of Trypanosoma cruzi tyrosine aminotransferase: substrate specificity is influenced by cofactor binding mode. Protein Sci.* **8**, 2406–2417.

Bochkarev, A., Barwell, J. A., Pfuetzner, R. A., Furey, W. Jr, Edwards, A. M. & Frappier, L. (1995). *Crystal structure of the DNA-binding domain of the Epstein–Barr virus origin-binding protein EBNA 1. Cell*, **83**, 39–46.

Bode, W., Mayr, I., Baumann, U., Huber, R., Stone, S. R. & Hofsteenge, J. (1989). *The refined 1.9 Å crystal structure of human alpha-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. EMBO J.* **8**, 3467–3475.

Bode, W., Reinemer, P., Huber, R., Kleine, T., Schnierer, S. & Tschesche, H. (1994). *The X-ray crystal structure of the catalytic domain of human neutrophil collagenase inhibited by a substrate analogue reveals the essentials for catalysis and specificity. EMBO J.* **13**, 1263–1269.

Bode, W., Wei, A. Z., Huber, R., Meyer, E., Travis, J. & Neumann, S. (1986). *X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. EMBO J.* **5**, 2453–2458.

Borkakoti, N., Winkler, F. K., Williams, D. H., D'Arcy, A., Broadhurst, M. J., Brown, P. A., Johnson, W. H. & Murray, E. J. (1994). *Structure of the catalytic domain of human fibroblast collagenase complexed with an inhibitor. Nature Struct. Biol.* **1**, 106–110.

Borst, P. (1999). *Multidrug resistance: a solvable problem? Ann. Oncol.* **10**, S162–S164.

Brandhuber, B. J., Boone, T., Kenney, W. C. & McKay, D. B. (1987). *Three-dimensional structure of interleukin-2. Science*, **238**, 1707–1709.

Brange, J. (1997). *The new era of biotech insulin analogues. Diabetologia*, **40**, S48–S53.

Breton, R., Housset, D., Mazza, C. & Fontecilla-Camps, J. C. (1996). *The structure of a complex of human 17-beta-hydroxysteroid dehydrogenase with estradiol and* $NADP^+$ *identifies two principal targets for the design of inhibitors. Structure*, **4**, 905–915.

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. & Wiley, D. C. (1993). *Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature (London)*, **364**, 33–39.

Browner, M. F., Smith, W. W. & Castelhano, A. L. (1995). *Matrilysin-inhibitor complexes: common themes among metalloproteases. Biochemistry*, **34**, 6602–6610.

Bruns, C. M., Hubatsch, I., Ridderstrom, M., Mannervik, B. & Tianer, J. A. (1999). *Human glutathione transferase A4-4 crystal structures and mutagenesis reveal the basis of high catalytic efficiency with toxic lipid peroxidation products. J. Mol. Biol.* **288**, 427–439.

Bruns, C. M., Nowalk, A. J., Arvai, A. S., McTigue, M. A., Vaughan, K. G., Mietzner, T. A. & McRee, D. E. (1997). *Structure of Hemophilus influenzae Fe(+3)-binding protein reveals convergent evolution within a superfamily. Nature Struct. Biol.* **4**, 919–924.

Brzozowski, A. M., Pike, A. C., Dauter, Z., Hubbard, R. E., Bonn, T., Engstrom, O., Ohman, L., Greene, G. L., Gustafsson, J. A. & Carlquist, M. (1997). *Molecular basis of agonism and antagonism in the oestrogen receptor. Nature (London)*, **389**, 753–758.

Bullock, T. L., Roberts, T. M. & Stewart, M. (1996). *2.5 Å resolution crystal structure of the motile major sperm protein (MSP) of Ascaris suum. J. Mol. Biol.* **263**, 284–296.

Burke, K. L., Dunn, G., Ferguson, M., Minor, P. D. & Almond, J. W. (1988). *Antigen chimaeras of poliovirus as potential new vaccines. Nature (London)*, **332**, 81–82.

Bussiere, D. E., Pratt, S. D., Katz, L., Severin, J. M., Holzman, T. & Park, C. H. (1998). *The structure of VanX reveals a novel amino-dipeptidase involved in mediating transposon-based vancomycin resistance. Mol. Cell*, **2**, 75–84.

Cameron, A. D., Sinning, I., L'Hermite, G., Olin, B., Board, P. G., Mannervik, B. & Jones, T. A. (1995). *Structural analysis of human alpha-class glutathione transferase A1-1 in the apo-form and in complexes with ethacrynic acid and its glutathione conjugate. Structure*, **3**, 717–727.

Carfi, A., Pares, S., Duee, E., Galleni, M., Duez, C., Frere, J. M. & Dideberg, O. (1995). *The 3-D structure of a zinc metallo-beta-lactamase from Bacillus cereus reveals a new type of protein fold. EMBO J.* **14**, 4914–4921.

Carrell, R. W. & Gooptu, B. (1998). *Conformational changes and diseases – serpins, prions and Alzheimer's. Curr. Opin. Struct. Biol.* **8**, 799–809.

Carrell, R. W., Stein, P. E., Fermi, G. & Wardell, M. R. (1994). *Biological implications of a 3 Å structure of dimeric antithrombin. Structure*, **2**, 257–270.

Carter, D. C. & Ho, J. X. (1994). *Structure of serum albumin. Adv. Protein Chem.* **45**, 153–203.

Cate, J. H., Yusupov, M. M., Zusupova, G. Z., Earnest, T. N. & Noller, H. F. (1999). *X-ray crystal structures of 70S ribosome functional complexes. Science*, **285**, 2095–2104.

Champness, J. N., Achari, A., Ballantine, S. P., Bryant, P. K., Delves, C. J. & Stammers, D. K. (1994). *The structure of Pneumocystis carinii dihydrofolate reductase to 1.9 Å resolution. Structure*, **2**, 915–924.

Chan, D. C., Fass, D., Berger, J. M. & Kim, P. S. (1997). *Core structure of gp41 from the HIV envelope glycoprotein. Cell*, **89**, 263–273.

Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. (1998). *Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. Science*, **282**, 2220–2226.

Chang, Y., Mochalkin, I., McCance, S. G., Cheng, B., Tulinsky, A. & Castellino, F. J. (1998). *Structure and ligand binding determinants of the recombinant kringle 5 domain of human plasminogen. Biochemistry*, **37**, 3258–3271.

Charifson, P. S. (1997). *Practical Application of Computer-Aided Drug Design*. New York: Marcel Dekker Inc.

Chitarra, V., Holm, I., Bentley, G. A., Petres, S. & Longacre, S. (1999). *The crystal structure of C-terminal merozoite surface protein 1 at 1.8 Å resolution, a highly protective malaria vaccine candidate. Mol. Cell*, **3**, 457–464.

Cho, Y., Gorina, S., Jeffrey, P. D. & Pavletich, N. P. (1994). *Crystal structure of a p53 tumor suppressor–DNA complex: understanding tumorigenic mutations. Science*, **265**, 346–355.

Choe, S., Bennett, M. J., Fujii, G., Curmi, P. M., Kantardjieff, K. A., Collier, R. J. & Eisenberg, D. (1992). *The crystal structure of diphtheria toxin. Nature*, **357**, 216–222.

Choi, H. J., Kang, S. W., Yang, C. H., Rhee, S. G. & Ryu, S. E. (1998). *Crystal structure of a novel human peroxidase enzyme at 2.0 Å resolution. Nature Struct. Biol.* **5**, 400–406.

Choudhury, D., Thompson, A., Stojanoff, V., Langermann, S., Pinkner, J., Hultgren, S. J. & Knight, S. D. (1999). *X-ray structure of the Fim-C-FimH chaperone–adhesin complex from uropathogenic Escherichia coli. Science*, **285**, 1061–1066.

Chudzik, D. M., Michels, P. A., de Walque, S. & Hol, W. G. J. (2000). *Structures of type 2 peroxisomal targeting signals in two trypanosomatid aldolases. J. Mol. Biol.* **300**, 697–707.

Cirilli, M., Zheng, R., Scapin, G. & Blanchard, J. S. (1993). *Structural symmetry: the three-dimensional structure of Haemophilus influenzae diaminopimelate epimerase. Biochemistry*, **37**, 16452–16458.

Ciszak, E. & Smith, G. D. (1994). *Crystallographic evidence for dual coordination around zinc in the T3R3 human insulin hexamer. Biochemistry*, **33**, 1512–1517.

Clackson, T., Yang, W., Rozamus, L. W., Hatada, M., Amara, J. F., Rollins, C. T., Stevenson, L. F., Magari, S. R., Wood, S. A., Courage, N. L., Lu, X., Cerasoli, F. J., Gilman, M. & Holt, D. A. (1998). *Redesigning an FKBP–ligand interface to generate chemical dimerizers with novel specificity. Proc. Natl Acad. Sci. USA*, **95**, 10437–10442.

Cleasby, A., Wonacott, A., Skarzynski, T., Hubbard, R. E., Davies, G. J., Proudfoot, A. E., Bernard, A. R., Payton, M. A. & Wells, T. N. (1996). *The X-ray crystal structure of phosphomannose isomerase from Candida albicans at 1.7 Å resolution. Nature Struct. Biol.* **3**, 470–479.

Clemons, W. M. J., May, J. L., Wimberly, B. T., McCutcheon, J. P., Capel, M. S. & Ramakrishnan, V. (1999). *Structure of a bacterial 30S ribosomal subunit at 5.5 Å resolution. Nature (London)*, **400**, 833–840.

Cobessi, D., Tete-Favier, F., Marchal, S., Azza, S., Branlant, G. & Aubry, A. (1999). *Apo and holo crystal structures of an NADP-dependent aldehyde dehydrogenase from Streptococcus mutans. J. Mol. Biol.* **290**, 161–173.

Colby, T. D., Vanderveen, K., Strickler, M. D., Markham, G. D. & Goldstein, B. M. (1999). *Crystal structure of human type II inosine monophosphate dehydrogenase: implications for ligand binding and drug design. Proc. Natl Acad. Sci. USA*, **96**, 3531–3536.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E. III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. & Barrell, B. G. (1998). *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature (London)*, **393**, 537–544.

Collins, F. S. (1992). *Cystic fibrosis: molecular biology and therapeutic implications. Science*, **256**, 774–779.

Concha, N. O., Rasmussen, B. A., Bush, K. & Herzberg, O. (1996). *Crystal structure of the wide-spectrum binuclear zinc beta-lactamase from Bacteroides fragilis. Structure*, **4**, 823–836.

Conlon, H. D., Baqai, J., Baker, K., Shen, Y. Q., Wong, B. L., Noiles, R. & Rausch, C. W. (1995). *2-step immobilized enzyme conversion of cephalosporin-c to 7-aminocephalosporanic acid. Biotechnol. Bioeng.* **46**, 510–513.

Cooper, J. B., McIntyre, K., Badasso, M. O., Wood, S. P., Zhang, Y., Garbe, T. R. & Young, D. (1995). *X-ray structure analysis of the iron-dependent superoxide dismutase from Mycobacterium tuberculosis at 2.0 Å resolution reveals novel dimer–dimer interactions. J. Mol. Biol.* **246**, 531–544.

Correll, C. C., Batie, C. J., Ballou, D. P. & Ludwig, M. L. (1992). *Phthalate dioxygenase reductase: a modular structure for electron transfer from pyridine nucleotides to [2Fe–2S]. Science*, **258**, 1604–1610.

Crennell, S., Garman, E., Laver, G., Vimr, E. & Taylor, G. (1994). *Crystal structure of Vibrio cholerae neuraminidase reveals dual lectin-like domains in addition to the catalytic domain. Structure*, **2**, 535–544.

Curry, S., Mandelkow, H., Brick, P. & Franks, N. (1998). *Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites. Nature Struct. Biol.* **5**, 827–835.

Cushman, D. W. & Ondetti, M. A. (1991). *History of the design of captopril and related inhibitors of angiotensin converting enzyme. Hypertension*, **17**, 589–592.

Cutfield, S. M., Dodson, E. J., Anderson, B. F., Moody, P. C., Marshall, C. J., Sullivan, P. A. & Cutfield, J. F. (1995). *The crystal structure of a major secreted aspartic proteinase from Candida albicans in complexes with two inhibitors. Structure*, **3**, 1261–1271.

Das, K., Ding, J., Hsiou, Y., Clark, A. D. Jr, Moereels, H., Koymans, L., Andries, K., Pauwels, R., Janssen, P. A. J., Boyer, P. L., Clark, P., Smith, R. H. Jr, Kroeger Smith, M. B., Michejda, C. J., Hughes, S. H. & Arnold, E. (1996). *Crystal structures of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. J. Mol. Biol.* **264**, 1085–1100.

Davies, J. F., Delcamp, T. J., Prendergast, N. J., Ashford, V. A., Freisheim, J. H. & Kraut, J. (1990). *Crystal structures of recombinant human dihydrofolate reductase complexed with folate and 5-deazafolate. Biochemistry*, **29**, 9467–9479.

De Bondt, H. L., Rosenblatt, J., Jancarik, J., Jones, H. D., Morgan, D. O. & Kim, S. H. (1993). *Crystal structure of cyclin-dependent kinase 2. Nature (London)*, **363**, 595–602.

Deller, M. C. & Jones, E. Y. (2000). *Cell surface receptors. Curr. Opin. Struct. Biol.* **10**, 213–219.

Derewenda, U., Derewenda, Z., Dodson, E. J., Dodson, G. G., Reynolds, C. D., Smith, G. D., Sparks, C. & Swenson, D. (1989). *Phenol stabilizes more helix in a new symmetrical zinc insulin hexamer. Nature (London)*, **338**, 594–596.

Dessen, A., Quemard, A., Blanchard, J. S., Jacobs, W. R. J. & Sacchettini, J. C. (1995). *Crystal structure and function of the isoniazid target of Mycobacterium tuberculosis. Science*, **267**, 1638–1641.

DeVos, A. M., Tong, L., Milburn, M. V., Matias, P. M., Jancarik, J., Noguchi, S., Nishimura, S., Miura, K., Ohtsuka, E. & Kim, S. H. (1988). *Three-dimensional structure of an oncogene protein: catalytic domain of human c-H-ras p21. Science*, **239**, 888–893.

DeVos, A. M., Ultsch, M. & Kossiakoff, A. A. (1992). *Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science*, **255**, 306–312.

Dhanaraj, V., Ye, Q.-Z., Johnson, L. L., Hupe, D. J., Otwine, D. F., Dunbar, J. B. J., Rubin, J. R., Pavlovsky, A., Humblet, C. & Blundell, T. L. (1996). *X-ray structure of a hydroxamate inhibitor complex of stromelysin catalytic domain and its comparison with members of the zinc metalloproteinase superfamily. Structure*, **4**, 375–386.

Dickerson, R. E. & Geis, I. (1983). *Hemoglobin.* Menlo Park: Benjamin Cummings Publishing Co.

Ding, J., Das, K., Tantillo, C., Zhang, W., Clark, A. D. Jr, Jessen, S., Lu, X., Hsiou, Y., Jacobo-Molina, A., Andries, K., Pauwels, R., Moereels, H., Koymans, L., Janssen, P. A. J., Smith, R. H. Jr, Kroeger Koepke, M., Michejda, C. J., Hughes, S. H. & Arnold, E. (1995). *Structure of HIV-1 reverse transcriptase in a complex with the non-nucleoside inhibitor alpha-APA R 95845 at 2.8 Å resolution. Structure*, **3**, 365–379.

Ding, J., McGrath, W. J., Sweet, R. M. & Mangel, W. F. (1996). *Crystal structure of the human adenovirus proteinase with its 11 amino acid cofactor. EMBO J.* **15**, 1778–1783.

Duggleby, H. J., Tolley, S. P., Hill, C. P., Dodson, E. J., Dodson, G. & Moody, P. C. (1995). *Penicillin acylase has a single-amino-acid catalytic centre. Nature (London)*, **373**, 264–268.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R. & Davies, D. R. (1994). *Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science*, **266**, 1981–1986.

Eads, J. C., Scapin, G., Xu, Y., Grubmeyer, C. & Sacchettini, J. C. (1994). *The crystal structure of human hypoxanthine-guanine phosphoribosyltransferase with bound GMP. Cell*, **78**, 325–334.

Ealick, S. E., Cook, W. J., Vijay-Kumar, S., Carson, M., Nagabhushan, T. L., Trotta, P. P. & Bugg, C. E. (1991). *Three-dimensional structure of recombinant human interferon-gamma. Science*, **252**, 698–702.

Ealick, S. E., Rule, S. A., Carter, D. C., Greenhough, T. J., Babu, Y. S., Cook, W. J., Habash, J., Helliwell, J. R., Stoeckler, J. D., Parks, R. E. J., Chen, S. F. & Bugg, C. E. (1990). *Three-dimensional structure of human* *erythrocytic purine nucleoside phosphorylase at 3.2-Å resolution. J. Biol. Chem.* **265**, 1812–1820.

Eckert, D. M., Malashkevish, V. N., Hong, L. H., Carr, P. A. & Kim, P. S. (1999). *Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket. Cell*, **99**, 103–115.

Ekstrom, J. L., Mathews, I. I., Stanley, B. A., Pegg, A. E. & Ealick, S. E. (1999). *The crystal structure of human S-adenosylmethionine decarboxylase at 2.25 Å resolution reveals a novel fold. Struct. Fold. Des.* **7**, 583–595.

Emsley, J., Cruz, M., Handin, R. & Liddington, R. (1998). *Crystal structure of the von Willebrand factor A1 domain and implications for the binding of platelet glycoprotein Ib. J. Biol. Chem.* **273**, 10396–10401.

Emsley, P., Charles, I. G., Fairweather, N. F. & Isaacs, N. W. (1996). *Structure of Bordetella pertussis virulence factor P.69 pertactin. Nature (London)*, **381**, 90–92.

Erickson, J., Neidhart, D. J., VanDrie, J., Kempf, D. J., Wang, X. C., Norbeck, D. W., Plattner, J. J., Rittenhouse, J. W., Turon, M., Wideburg, N., Kohlbrenner, W. E., Simmer, R., Helfrich, R., Paul, D. A. & Knigge, M. (1990). *Design, activity, and 2.8 Å crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. Science*, **249**, 527–533.

Erickson, J. W. & Burt, S. K. (1996). *Structural mechanisms of HIV drug resistance. Annu. Rev. Pharmacol. Toxicol.* **36**, 545–571.

Erlandsen, H., Fusetti, F., Martinez, A., Hough, E., Flatmark, T. & Stevens, R. C. (1997). *Crystal structure of the catalytic domain of human phenylalanine hydroxylase reveals the structural basis for phenylketonuria. Nature Struct. Biol.* **4**, 995–1000.

Esser, C. K., Bugianesi, R. L., Caldwell, C. G., Chapman, K. T., Durette, P. L., Girotra, N. N., Kopka, I. E., Lanza, T. J., Levorse, D. A., Maccoss, M., Owens, K. A., Ponpipom, M. M., Simeone, J. P., Harrison, R. K., Niedzwiecki, L., Becker, J. W., Marcy, A. I., Axel, M. G., Christen, A. J., McDonnell, J., Moore, V. L., Olsqewski, J. M., Saphos, C., Visco, D. M., Shen, F., Colletti, A., Kriter, P. A. & Hagmann, W. K. (1997). *Inhibition of stromelysin-1 (MMP-3) by P1′-biphenylylethyl carboxyalkyl dipeptides. J. Med. Chem.* **40**, 1026–1040.

Fan, C., Moews, P. C., Walsh, C. T. & Knox, J. R. (1994). *Vancomycin resistance: structure of D-alanine:D-alanine ligase at 2.3 Å resolution. Science*, **266**, 439–443.

Fan, E., Zhang, Z., Minke, W. E., Hou, Z., Verlinde, C. L. M. J. & Hol, W. G. J. (2000). *A 10^5 gain in affinity for pentavalent ligands of E. coli heat-labile enterotoxin by modular structure-based design. J. Am. Chem. Soc.* **122**, 2663–2664.

Ferrer, M., Kapoor, T. M., Strassmaier, T., Weissenhorn, W., Skehel, J. J., Oprian, D., Schreiber, S. L., Wiley, D. C. & Harrison, S. C. (1999). *Selection of gp41-mediated HIV-1 cell entry inhibitors from biased combinatorial libraries of non-natural binding elements. Nature Struct. Biol.* **6**, 953–960.

Fields, B. A., Malchiodi, E. L., Li, H., Ysern, X., Stauffacher, C. V., Schlievert, P. M., Karjalainen, K. & Mariuzza, R. A. (1996). *Crystal structure of a T-cell receptor beta-chain complexed with a superantigen. Nature (London)*, **384**, 188–192.

Filman, D. J., Wien, M. W., Cunningham, J. A., Bergelson, J. M. & Hogle, J. M. (1998). *Structure determination of echovirus 1. Acta Cryst.* D**54**, 1261–1272.

Finzel, B. C., Baldwin, E. T., Bryant, G. L. J., Hess, G. F., Wilks, J. W., Trepod, C. M., Mott, J. E., Marshall, V. P., Petzold, G. L., Poorman, R. A., O'Sullivan, T. J., Schostarez, H. J. & Mitchell, M. A. (1998). *Structural characterizations of nonpeptidic thiadiazole inhibitors of matrix metalloproteinases reveal the basis for stromelysin selectivity. Protein Sci.* **7**, 2118–2126.

Focia, P. J., Craig, S. P. III, Nieves-Alicea, R., Fletterick, R. J. & Eakin, A. E. (1998). *A 1.4 Å crystal structure for the hypoxanthine phosphoribosyltransferase of Trypanosoma cruzi. Biochemistry*, **37**, 15066–15075.

Foster, B. A., Coffey, H. A., Morin, M. J. & Rastinejad, F. (1999). *Pharmacological rescue of mutant p53 conformation and function. Science*, **286**, 2507–2510.

Fox, M. P., Otto, M. J. & McKinlay, M. A. (1986). *Prevention of rhinovirus and poliovirus uncoating by WIN 51711, a new antiviral drug. Antimicrob. Agents Chemother.* **30**, 110–116.

Frankenberg, N., Erskine, P. T., Cooper, J. B., Shoolingin-Jordan, P. M., Jahn, D. & Heinz, D. W. (1999). *High resolution crystal structure of a Mg^{2+}-dependent porphobilinogen synthase. J. Mol. Biol.* **289**, 591–602.

Fremont, D. H., Matsumura, M., Stura, E. A., Peterson, P. A. & Wilson, I. A. (1992). *Crystal structures of two viral peptides in complex with murine MHC class I H-2Kb. Science*, **257**, 919–927.

Freymann, D., Down, J., Carrington, M., Roditi, I., Turner, M. & Wiley, D. (1990). *2.9 Å resolution structure of the N-terminal domain of a variant surface glycoprotein from Trypanosoma brucei. J. Mol. Biol.* **216**, 141–160.

Fulop, V., Ridout, C. J., Greenwood, C. & Hajdu, J. (1995). *Crystal structure of the di-haem cytochrome c peroxidase from Pseudomonas aeruginosa. Structure*, **3**, 1225–1233.

Futterer, K., Wong, J., Grucza, R. A., Chan, A. C. & Waksman, G. (1998). *Structural basis for Syk tyrosine kinase ubiquity in signal transduction pathways revealed by the crystal structure of its regulatory SH2 domains bound to a dually phosphorylated ITAM peptide. J. Mol. Biol.* **281**, 523–537.

Gaboriaud, C., Serre, L., Guy-Crotte, O., Forest, E. & Fontecilla-Camps, J. C. (1996). *Crystal structure of human trypsin 1: unexpected phosphorylation of Tyr151. J. Mol. Biol.* **259**, 995–1010.

Gamblin, S. J., Cooper, B., Millar, J. R., Davies, G. J., Littlechild, J. A. & Watson, H. C. (1990). *The crystal structure of human muscle aldolase at 3.0 Å resolution. FEBS Lett.* **264**, 282–286.

Garboczi, D. N., Ghosh, P., Utz, U., Fan, Q. R., Biddison, W. E. & Wiley, D. C. (1996). *Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature (London)*, **384**, 134–141.

Garcia, K. C., Degano, M., Stanfield, R. L., Brunmark, A., Jackson, M. R., Petereson, P. A., Teyton, L. & Wilson, I. A. (1996). *An $\alpha\beta$ T cell receptor structure at 2.5 Å and its orientation in the TCR–MHC complex. Science*, **274**, 209–219.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D. C., Pertea, M., Salzburg, S., Zhou, L., Sutton, G. G., Clayton, R., White, O., Smith, H. O., Fraser, C. M., Adams, M. D., Venter, J. C. & Hoffman, S. L. (1998). *Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science*, **282**, 1126–1132.

Gatti, D. L., Palfey, B. A., Lah, M. S., Entsch, B., Massey, V., Ballou, D. P. & Ludwig, M. L. (1994). *The mobile flavin of 4-OH benzoate hydroxylase. Science*, **266**, 110–114.

Ghosh, D., Pletnev, V. Z., Zhy, D. W., Wawrzak, Z., Duax, W. L., Pangborn, W., Labrie, F. & Lin, S. W. (1995). *Structure of human estrogenic 17-beta-hydroxysteroid dehydrogenase at 2.20-Å resolution. Structure*, **3**, 503–513.

Giulian, D., Corpuz, M., Richmond, B., Wendt, E. & Hall, E. R. (1996). *Activated microglia are the principal glial source of thromboxane in the central nervous system. Neurochem. Int.* **29**, 65–76.

Gohlke, U., Gomis-Ruth, F. X., Crabbe, T., Murphy, G., Docherty, A. J. & Bode, W. (1996). *The C-terminal (haemopexin-like) domain structure of human gelatinase A (MMP2): structural implications for its function. FEBS Lett.* **378**, 126–120.

Gomis-Ruth, F. X., Gohlke, U., Betz, M., Knauper, V., Murphy, G., Lopez-Otin, C. & Bode, W. (1996). *The helping hand of collagenase-3 (MMP-13): 2.7 Å crystal structure of its C-terminal haemopexin-like domain. J. Mol. Biol.* **264**, 556–566.

Gomis-Ruth, F. X., Maskos, K., Betz, M., Bergner, A., Huber, R., Suzuki, K., Yoshida, N., Nagase, H., Brew, K., Bourenkov, G. P., Bartunik, H. & Bode, W. (1997). *Mechanism of inhibition of the human matrix metalloproteinase stromelysin-1 by TIMP-1. Nature (London)*, **389**, 77–81.

Gong, W., O'Gara, M., Blumenthal, R. M. & Cheng, X. (1997). *Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res.* **25**, 2702–2715.

Goodwill, K. E., Sabatier, C., Marks, C., Raag, R., Fitzpatrick, P. F. & Stevens, R. C. (1997). *Crystal structure of tyrosine hydroxylase at 2.3 Å and its implications for inherited neurodegenerative diseases. Nature Struct. Biol.* **4**, 578–585.

Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). *Energy functions for protein design. Curr. Opin. Struct. Biol.* **9**, 509–513.

Gorina, S. & Pavletich, N. P. (1996). *Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. Science*, **274**, 1001–1005.

Gouet, P., Jouve, H. M. & Dideberg, O. (1995). *Crystal structure of Proteus mirabilis PT catalase with and without bound NADPH. J. Mol. Biol.* **249**, 933–954.

Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R., Isaacs, N. W. & Sawyer, L. (1999). *The two types of 3-dehydrogenase have distinct structures but catalyze the same overall reaction. Nature Struct. Biol.* **6**, 521–525.

Grant, R. A., Hiemath, C. N., Filman, D. J., Syed, R., Andries, K. & Hogle, J. M. (1994). *Structures of poliovirus complexes with anti-viral drugs: implications for viral stability and drug design. Curr. Biol.* **4**, 784–797.

Graves, B. J., Hatada, M. H., Hendrickson, W. A., Miller, J. K., Madison, V. S. & Satow, Y. (1990). *Structure of interleukin 1 alpha at 2.7-Å resolution. Biochemistry*, **29**, 2679–2684.

Grimes, J., Basak, A. K., Roy, P. & Stuart, D. (1995). *The crystal structure of bluetongue virus VP7. Nature (London)*, **373**, 167–170.

Hampele, I. C., D'Arcy, A., Dale, G. E., Kostrewa, D., Nielsen, J., Oefner, C., Page, M. G., Schonfeld, H. J., Stuber, D. & Then, R. L. (1997). *Structure and function of the dihydropteroate synthase from Staphylococcus aureus. J. Mol. Biol.* **268**, 21–30.

Han, S., Eltis, L. D., Timmis, K. N., Muchmore, S. W. & Bolin, J. T. (1995). *Crystal structure of the biphenyl-cleaving extradiol dioxygenase from a PCB-degrading pseudomonad. Science*, **270**, 976–980.

Hansen, J. L., Long, A. M. & Schultz, S. C. (1997). *Structure of the RNA-dependent RNA polymerase of poliovirus. Structure*, **5**, 1109–1122.

Harrington, D. J., Adachi, K. & Royer, W. E. Jr (1997). *The high resolution crystal structure of deoxyhemoglobin S. J. Mol. Biol.* **272**, 398–407.

Harris, S. F. & Botchan, M. R. (1999). *Crystal structure of the human papillomavirus type 18 E2 activation domain. Science*, **284**, 1673–1677.

He, X. M. & Carter, D. C. (1992). *Atomic structure and chemistry of human serum albumin. Nature (London)*, **358**, 209–215.

Hegde, R. S. & Androphy, E. J. (1998). *Crystal structure of the E2 DNA-binding domain from human papillomavirus type 16: implications for its DNA binding-site selection mechanism. J. Mol. Biol.* **284**, 1479–1489.

Hennig, M., Dale, G. E., D'Arcy, A., Danel, F., Fischer, S., Gray, C. P., Jolidon, S., Muller, F., Page, M. G., Pattison, P. & Oefner, C. (1999). *The structure and function of the 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase from Haemophilus influenzae. J. Mol. Biol.* **287**, 211–219.

Hennig, M., D'Arcy, A., Hampele, I. C., Page, M. G., Oefner, C. & Dale, G. E. (1998). *Crystal structure and reaction mechanism of 7,8-dihydroneopterin aldolase from Staphylococcus aureus. Nature Struct. Biol.* **5**, 357–362.

Herzberg, O. & Moult, J. (1987). *Bacterial resistance to beta-lactam antibiotics: crystal structure of beta-lactamase from Staphylococcus aureus PC1 at 2.5 Å resolution. Science*, **236**, 694–701.

Hill, C. P., Worthylake, D., Bancroft, D. P., Christensen, A. M. & Sundquist, W. I. (1996). *Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. Proc. Natl Acad. Sci. USA*, **93**, 3099–3104.

Hodel, A. E., Gershon, P. D., Shi, X. & Quiocho, F. A. (1996). *The 1.85 Å structure of vaccinia protein VP39: a bifunctional enzyme that participates in the modification of both mRNA ends. Cell*, **85**, 247–256.

Hodgkin, D. C. (1971). *Insulin molecules: the extent of our knowledge. Pure Appl. Chem.* **26**, 375–384.

Hofmann, B., Schomburg, D. & Hecht, H. J. (1993). *Crystal structure of a thiol proteinase from Staphylococcus aureus V-8 in the E-64 inhibitor complex. Acta Cryst.* A**49**, C-102.

Hogle, J. M., Chow, M. & Filman, D. J. (1985). *Three-dimensional structure of poliovirus at 2.9 Å resolution. Science*, **229**, 1358–1365.

Hol, W. G. J. (1986). *Protein crystallography and computer graphics – toward rational drug design. Angew. Chem. Int. Ed. Engl.* **25**, 767–778.

Hoog, S. S., Smith, W. W., Qiu, X., Janson, C. A., Hellmig, B., McQueney, M. S., O'Donnell, K., O'Shannessy, D., DiLella, A. G., Debouck, C. & Abdel-Meguid, S. S. (1997). *Active site cavity of herpesvirus proteases revealed by the crystal structure of herpes simplex virus protease/inhibitor complex. Biochemistry*, **36**, 14023–14029.

Hough, E., Hansen, L. K., Birknes, B., Jynge, K., Hansen, S., Hordvik, A., Little, C., Dodson, E. & Derewenda, Z. (1989). *High-resolution (1.5 Å) crystal structure of phospholipase C from Bacillus cereus. Nature (London)*, **338**, 357–360.

Hsiou, Y., Das, K., Ding, J., Clark, A. D. Jr, Kleim, J. P., Rosner, M., Winkler, I., Riess, G., Hughes, S. H. & Arnold, E. (1998). *Structures of Tyr188Leu mutant and wild-type HIV-1 reverse transcriptase complexed with the non-nucleoside inhibitor HBY 097: inhibitor flexibility is a useful design feature for reducing drug resistance. J. Mol. Biol.* **284**, 313–323.

Hu, S. H., Peek, J. A., Rattigan, E., Taylor, R. K. & Martin, J. L. (1997). *Structure of TcpG, the DsbA protein folding catalyst from Vibrio cholerae. J. Mol. Biol.* **268**, 137–146.

Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. (1998). *Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. Science*, **282**, 1669–1675.

Huang, K., Strynadka, N. C., Bernard, V. D., Peanasky, R. J. & James, M. N. (1994). *The molecular structure of the complex of Ascaris chymotrypsin/elastase inhibitor with porcine elastase. Structure*, **2**, 679–689.

Huang, S., Xue, Y., Sauer-Eriksson, E., Chirica, L., Lindskog, S. & Jonsson, B. H. (1998). *Crystal structure of carbonic anhydrase from Neisseria gonorrhoeae and its complex with the inhibitor acetazolamide. J. Mol. Biol.* **283**, 301–310.

Hubbard, S. R. (1997). *Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. EMBO J.* **16**, 5572–5581.

Hubbard, S. R., Wei, L., Ellis, L. & Hendrickson, W. A. (1994). *Crystal structure of the tyrosine kinase domain of the human insulin receptor. Nature (London)*, **372**, 746–754.

Huizinga, E. G., Martijn van der Plas, R., Kroon, J., Sixma, J. J. & Gros, P. (1997). *Crystal structure of the A3 domain of human von Willebrand factor: implications for collagen binding. Structure*, **5**, 1147–1156.

Hulsmeyer, M., Hecht, H. J., Niefind, K., Hofer, B., Eltis, L. D., Timmis, K. N. & Schomburg, D. (1998). *Crystal structure of cis-biphenyl-2,3-dihydrodiol-2,3-dehydrogenase from a PCB degrader at 2.0 Å resolution. Protein Sci.* **7**, 1286–1293.

Hurley, T. D., Bosron, W. F., Hamilton, J. A. & Amzel, L. M. (1991). *Structure of human beta 1 beta 1 alcohol dehydrogenase: catalytic effects of non-active-site substitutions. Proc. Natl Acad. Sci. USA*, **88**, 8149–8153.

Isupov, M. N., Antson, A. A., Dodson, E. J., Dodson, G. G., Dementieva, I. S., Zakomirdina, L. N., Wilson, K. S., Dauter, Z., Lebedev, A. A. & Harutyunyan, E. H. (1998). *Crystal structure of tryptophanase. J. Mol. Biol.* **276**, 603–623.

Itzstein, M. von, Wu, W. Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., Van Phan, T., Smythe, M. L., White, H. F., Oliver, S. W., Colman, P. M., Varghese, J. N., Ryan, D. M., Woods, J. M., Bethell, R. C., Hotham, V. J., Cameron, J. M. & Penn, C. R. (1993). *Rational design of potent sialidase-based inhibitors of influenza virus replication. Nature (London)*, **363**, 418–423.

Jackson, R. C. (1997). *Contributions of protein structure-based drug design to cancer chemotherapy. Semin. Oncol.* **24**, 164–172.

Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark, A. D. Jr, Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H. & Arnold, E. (1993). *Crystal structure of human immunodeficiency virus type 1 reverse transcriptase complexed with double-stranded DNA at 3.0 Å resolution shows bent DNA. Proc. Natl Acad. Sci. USA*, **90**, 6320–6324.

Jain, S., Drendel, W. B., Chen, Z. W., Mathews, F. S., Sly, W. S. & Grubb, J. H. (1996). *Structure of human beta-glucuronidase reveals candidate lysosomal targeting and active-site motifs. Nature Struct. Biol.* **3**, 375–381.

Jia, Z., Vandonselaar, M., Quail, J. W. & Delbaere, L. T. (1993). *Active-centre torsion-angle strain revealed in 1.6 Å-resolution structure of histidine-containing phosphocarrier protein. Nature (London)*, **361**, 94–97.

Kallarakal, A. T., Mitra, B., Kozarich, J. W., Gerlt, J. A., Clifton, J. G., Petsko, G. A. & Kenyon, G. L. (1995). *Mechanism of the reaction catalyzed by mandelate racemase: structure and mechanistic properties of the K166R mutant. Biochemistry*, **34**, 2788–2797.

Kallen, J., Spitzfaden, C., Zurini, M. G. M., Wider, G., Widmer, H., Wuethrich, K. & Walkinshaw, M. D. (1991). *Structure of human cyclophilin and its binding site for cyclosporin A determined by X-ray crystallography and NMR spectroscopy. Nature (London)*, **353**, 276–279.

Kannan, K. K., Notstrand, B., Fridborg, K., Lovgren, S., Ohlsson, A. & Petef, M. (1975). *Crystal structure of human erythrocyte carbonic anhydrase B. Three-dimensional structure at a nominal 2.2-Å resolution. Proc. Natl Acad. Sci. USA*, **72**, 51–55.

Karpusas, M., Nolte, M., Benton, C. B., Meier, W., Lipscomb, W. N. & Goelz, S. (1997). *The crystal structure of human interferon beta at 2.2-Å resolution. Proc. Natl Acad. Sci. USA*, **94**, 11813–11818.

Ke, H., Zydowsky, L. D., Liu, J. & Walsh, C. T. (1991). *Crystal structure of recombinant human T-cell cyclophilin A at 2.5-Å resolution. Proc. Natl Acad. Sci. USA*, **88**, 9483–9487.

Kim, H., Certa, U., Döbelli, H., Jakob, P. & Hol, W. G. J. (1998). *Crystal structure of fructose-1,6-bisphosphate aldolase from the human malaria parasite Plasmodium falciparum. Biochemistry*, **37**, 4388–4396.

Kim, H., Feil, I. K., Verlinde, C. L. M. J., Petra, P. H. & Hol, W. G. J. (1995). *Crystal structure of glycosomal glyceraldehyde-3-phosphate dehydrogenase from Leishmania mexicana: implication for structure-based drug design and a new position for the inorganic phosphate binding site. Biochemistry*, **34**, 14975–14986.

Kim, K. K., Song, H. K., Shin, D. H., Hwang, K. Y. & Suh, S. W. (1997). *The crystal structure of a triacylglycerol lipase from Pseudomonas cepacia reveals a highly open conformation in the absence of a bound inhibitor. Structure*, **5**, 173–185.

Kim, Y., Yoon, K.-H., Khang, Y., Turley, S. & Hol, W. G. J. (2000). *The 2.0 Å crystal structure of cephalosporin acylase. Struct. Fold. Des.* **8**, 1059–1068.

Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A., Kalish, V. J., Tucker, K. D., Showalter, R. E., Moomaw, E. W., Gastinel, L. N., Habuka, N., Chen, X., Maldonado, F., Barker, J. E., Bacquet, R. & Villafranca, J. E. (1995). *Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. Nature (London)*, **378**, 641–644.

Kitadokoro, K., Hagishita, S., Sato, T., Ohtan, M. & Miki, K. (1998). *Crystal structure of human secretory phospholipase A2-IIA complex with the potent indolizine inhibitor 120–1032. J. Biochem.* **123**, 619–623.

Kitov, P. I., Sadowska, J. M., Mulvey, G., Armstrong, G. D., Ling, H., Pannu, N. S., Read, R. J. & Bundle, D. R. (2000). *Shiga-like toxins are neutralized by tailored multivalent carbohydrate ligands. Nature (London)*, **403**, 669–672.

Knighton, D. R., Kan, C. C., Howland, E., Janson, C. A., Hostomska, Z., Welsh, K. M. & Matthews, D. A. (1994). *Structure of and kinetic channelling in bifunctional dihydrofolate reductase-thymidylate synthase. Nature Struct. Biol.* **1**, 186–194.

Ko, T.-P., Safo, M. K., Musayev, F. N., Di Salvo, M. L., Wang, C., Wu, S.-H. & Abraham, D. J. (2000). *Structure of human erythrocyte catalase. Acta Cryst.* D**56**, 241–245.

Kohara, M., Abe, S., Komatsu, T., Tago, K., Arita, M. & Nomoto, A. (1988). *A recombinant virus between the Sabin 1 and Sabin 3 vaccine strains of poliovirus as a possible candidate for a new type 3 poliovirus live vaccine strain. J. Virol.* **62**, 2828–2835.

Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. & Steitz, T. A. (1992). *Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science*, **256**, 1783–1790.

Kohno, M., Funatsu, J., Mikami, B., Kugimiya, W., Matsuo, T. & Morita, Y. (1996). *The crystal structure of lipase II from Rhizopus niveus at 2.2 Å resolution. J. Biochem.* **120**, 505–510.

Kopka, M. L., Yoon, C., Goodsell, D., Pjura, P. & Dickerson, R. E. (1985). *The molecular origin of DNA-drug specificity in netropsin and distamycin. Proc. Natl Acad. Sci. USA*, **82**, 1376–1380.

Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W. & Wittinghofer, A. (1990). *Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35-Å resolution: implications for the mechanism of GTP hydrolysis. EMBO J.* **9**, 2351–2359.

Kuntz, I. D. (1992). *Structure-based strategies for drug design and discovery. Science*, **257**, 1078–1082.

Kurihara, H., Mitsui, Y., Ohgi, K., Irie, M., Mizuno, H. & Nakamura, K. T. (1992). *Crystal and molecular structure of RNase Rh, a new class of microbial ribonuclease from Rhizopus niveus. FEBS Lett.* **306**, 189–192.

Kuzin, A. P., Nukaga, M., Nukaga, Y., Hujer, A. M., Bonomo, R. A. & Knox, J. R. (1999). *Structure of the SHV-1 beta-lactamase. Biochemistry*, **38**, 5720–5727.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein incomplex with the CD4 receptor and a neutralizing human antibody. Nature (London)*, **393**, 648–659.

Laba, D., Bauer, M., Huber, R., Fischer, S., Rudolph, R., Kohnert, U. & Bode, W. (1996). *The 2.3 Å crystal structure of the catalytic domain of recombinant two-chain human tissue-type plasminogen activator. J. Mol. Biol.* **258**, 117–135.

Lacy, D. B. & Stevens, R. C. (1998). *Unraveling the structure and modes of action of bacterial toxins. Curr. Opin. Struct. Biol.* **8**, 778–784.

Lacy, D. B., Tepp, W., Cohen, A. C., DasGupta, B. R. & Stevens, R. C. (1998). *Crystal structure of botulinum neurotoxin type A and implications for toxicity. Nature Struct. Biol.* **5**, 898–902.

Lantwin, C. B., Schlichting, I., Kabsch, W., Pai, E. F. & Krauth-Siegel, R. L. (1994). *The structure of Trypanosoma cruzi trypanothione reductase in the oxidized and NADPH reduced state. Proteins*, **18**, 161–173.

Le, H. V., Yao, N. & Weber, P. C. (1998). *Emerging targets in the treatment of hepatitis C infection. Emerg. Ther. Targets*, **2**, 125–136.

Lebron, J. A., Bennett, M. J., Vaughn, D. E., Chirino, A. J., Snow, P. M., Mintier, G. A., Feder, J. N. & Bjorkman, P. J. (1998). *Crystal structure of the hemochromatosis protein HFE and characterization of its interaction with transferrin receptor. Cell*, **93**, 111–123.

Lee, C. H., Saksela, K., Mirza, U. A., Chait, B. T. & Kuriyan, J. (1996). *Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. Cell*, **85**, 931–942.

Leonard, S. A., Gittis, A. G., Petrella, E. D., Pollard, T. D. & Lattman, E. E. (1997). *Crystal structure of the actin-binding protein actophorin from Acanthamoeba. Nature Struct. Biol.* **4**, 369–373.

Levine, M. M. & Noriega, F. (1995). *A review of the current status of enteric vaccines. Papua New Guinea Med. J.* **38**, 325–331.

Li, H., Dunn, J. J., Luft, B. J. & Lawson, C. L. (1997). *Crystal structure of Lyme disease antigen outer surface protein A complexed with an Fab. Proc. Natl Acad. Sci. USA*, **94**, 3584–3589.

Li, R., Sirawaraporn, P., Chitnumsub, P., Sirawaraporn, W. & Hol, W. G. J. (2000). *Three-dimensional structure of M. tuberculosis dihydrofolate reductase reveals opportunities for the design of novel tuberculosis drugs. J. Mol. Biol.* **295**, 307–323.

Li de la Sierra, I., Pernot, L., Prange, T., Saludjian, P., Schiltz, M., Fourme, R. & Padron, G. (1997). *Molecular structure of the lipoamide dehydrogenase domain of a surface antigen from Neisseria meningitidis. J. Mol. Biol.* **269**, 129–141.

Libson, A. M., Gittis, A. G., Collier, I. E., Marmer, B. L., Goldberg, G. I. & Lattman, E. E. (1995). *Crystal structure of the haemopexin-like C-terminal domain of gelatinase A. Nature Struct. Biol.* **2**, 938–942.

Liljas, A., Kannan, K. K., Bergsten, P. C., Waara, I., Fridborg, K., Strandberg, B., Carlbom, U., Jarup, L., Lovgren, S. & Petef, M. (1972). *Crystal structure of human carbonic anhydrase C. Nat. New Biol.* **235**, 131–137.

Lin, J. H., Ostovic, D. & Vacca, J. P. (1998). *The integration of medicinal chemistry, drug metabolism, and pharmaceutical research and development in drug discovery and development. The story of Crixivan, an HIV protease inhibitor. Pharm. Biotechnol.* **99**, 233–255.

Ling, H., Boodhoo, A., Hazes, B., Cummings, M. D., Armstrong, G. D., Brunton, J. L. & Read, R. J. (1998). *Structure of the shiga-like toxin I B-pentamer complexed with an analogue of its receptor Gb3. Biochemistry*, **37**, 1777–1788.

Liu, S., Fedorov, A. A., Pollard, T. D., Lattman, E. E., Almo, S. C. & Magnus, K. A. (1998). *Crystal packing induces a conformational change in profilin-I from Acanthamoeba castellanii. J. Struct. Biol.* **123**, 22–29.

Liu, Y., Gong, W., Huang, C. C., Herr, W. & Cheng, X. (1999). *Crystal structure of the conserved core of the herpes simplex virus transcriptional regulatory protein VP16. Genes Dev.* **13**, 1692–1703.

Livnah, O., Johnson, D. L., Stura, E. A., Farrell, F. X., Barbone, F. P., You, Y., Liu, K. D., Goldsmith, M. A., He, W., Krause, C. D., Petska, S., Jolliffe, L. K. & Wilson, I. A. (1998). *An antagonist peptide-EPO receptor complex suggests that receptor dimerization is not sufficient for activation. Nature Struct. Biol.* **5**, 993–1004.

Livnah, O., Stura, E. A., Johnson, D. L., Middleton, S. A., Mulcahy, L. S., Wrighton, N. C., Dower, W. J., Jolliffe, L. K. & Wilson, I. A. (1996). *Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å. Science*, **273**, 464–471.

Lobkovsky, E., Moews, P. C., Liu, H., Zhao, H., Frere, J. M. & Knox, J. R. (1993). *Evolution of an enzyme activity: crystallographic structure at 2-Å resolution of cephalosporinase from the ampC gene of Enterobacter cloacae P99 and comparison with a class A penicillinase. Proc. Natl Acad. Sci. USA*, **90**, 11257–11261.

Loebermann, H., Tokuoka, R., Deisenhofer, J. & Huber, R. (1984). *Human alpha 1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function. J. Mol. Biol.* **177**, 531–557.

Loll, P. J. & Lattman, E. E. (1989). *The crystal structure of the ternary complex of staphylococcal nuclease, Ca^{2+}, and the inhibitor pdTp, refined at 1.65 Å. Proteins*, **5**, 183–201.

Love, R. A., Parge, H. E., Wickersham, J. A., Hostomsky, Z., Habuka, N., Moomaw, E. W., Adachi, T. & Hostomska, Z. (1996). *The crystal structure of hepatitis C virus NS3 proteinase reveals a trypsin-like fold and a structural zinc binding site. Cell*, **87**, 331–342.

Lovejoy, B., Cleasby, A., Hassell, A. M., Longley, K., Luther, M. A. W. D., McGeehan, G., McElroy, A. B., Drewry, D., Lambert, M. H. & Jordan, S. R. (1994). *Structure of the catalytic domain of fibroblast collagenase complexed with an inhibitor. Science*, **263**, 375–377.

Lovejoy, B., Hassell, A. M., Luther, M. A., Weigl, D. & Jordan, S. R. (1994). *Crystal structures of recombinant 19-kDa human fibroblast collagenase complexed to itself. Biochemistry*, **33**, 8207–8217.

Lukatela, G., Krauss, N., Theis, K., Selmer, T., Gieselmann, V., von Figura, K. & Saenger, W. (1998). *Crystal structure of human arylsulfatase A: the aldehyde function and the metal ion at the active site suggest a novel mechanism for sulfate ester hydrolysis. Biochemistry*, **37**, 3654–3664.

McGrath, M. E., Eakin, A. E., Engel, J. C., McKerrow, J. H., Craik, C. S. & Fletterick, R. J. (1995). *The crystal structure of cruzain: a therapeutic target for Chagas' disease. J. Mol. Biol.* **247**, 251–259.

McGrath, M. E., Palmer, J. T., Bromme, D. & Somoza, J. R. (1998). *Crystal structure of human cathepsin S. Protein Sci.* **7**, 1294–1302.

McLaughlin, P. J., Gooch, J. T., Mannherz, H. G. & Weeds, A. G. (1993). *Structure of gelsolin segment 1-actin complex and the mechanism of filament severing. Nature (London)*, **364**, 685–692.

McTigue, M. A., Wickersham, J. A., Pinko, C., Showalter, R. E., Parast, C. V., Tempczyk-Russell, A., Gehring, M. R., Mroczkowski, B., Kan, C. C., Villafranca, J. E. & Appelt, K. (1999). *Crystal structure of the kinase domain of human vascular endothelial growth factor receptor 2: a key enzyme in angiogenesis. Struct. Fold. Des.* **7**, 319–330.

McTigue, M. A., Williams, D. R. & Tainer, J. A. (1995). *Crystal structure of a schistosomal drug and vaccine target: glutathione S-transferase from Schistosoma japonica and its complex with the leading antischistosomal drug praziquantel. J. Mol. Biol.* **246**, 21–27.

Maldonado, E., Soriano-Garcia, M., Moreno, A., Cabrera, N., Garza-Ramos, G., de Gomez-Puyou, M., Gomez-Puyou, A. & Perez-Montfort, R. (1998). *Differences in the intersubunit contacts in triosephosphate isomerase from two closely related pathogenic trypanosomes. J. Mol. Biol.* **283**, 193–203.

Malik, P. & Perham, R. N. (1997). *Simultaneous display of different peptides on the surface of filamentous bacteriophage. Nucleic Acids Res.* **25**, 915–916.

Mande, S. C., Mainfroid, V., Kalk, K. H., Goraj, K., Marital, J. A. & Hol, W. G. J. (1994). *Crystal structure of recombinant human triosephosphate isomerase at 2.8 Å resolution. Protein Sci.* **3**, 810–821.

Mande, S. C., Mehra, F., Bloom, B. R. & Hol, W. G. J. (1996). *Structure of the heat shock protein chaperonin-10 of Mycobacterium leprae. Science*, **271**, 203–207.

Martin, A., Wychowski, C., Couderc, T., Crainic, R., Hogle, J. & Girard, M. (1988). *Engineering a poliovirus type 2 antigenic site on a type 1 capsid results in a chimaeric virus which is neurovirulent for mice. EMBO J.* **7**, 2839–2847.

Mather, T., Oganesseyan, V., Hof, P., Huber, R., Foundling, S., Esmon, S. & Bode, W. (1996). *The 2.8 Å crystal structure of Gla-domainless activated protein C. EMBO J.* **15**, 6822–6831.

Mathews, I. I., Vanderhoff-Hanaver, P., Castellino, F. J. & Tulinsky, A. (1996). *Crystal structures of the recombinant kringle 1 domain of human plasminogen in complexes with the ligands epsilon-aminocaproic acid and trans-4-(aminomethyl)cyclohexane-1-carboxylic acid. Biochemistry*, **35**, 2567–2576.

Matthews, D. A., Alden, R. A., Bolin, J. T., Freer, S. T., Hamlin, R., Xuong, N., Kraut, J., Poe, M., Williams, M. & Hoogsteen, K. (1977). *Dihydrofolate reductase: X-ray structure of the binary complex with methotrexate. Science*, **197**, 452–455.

Matthews, D. A., Smith, W. W., Ferre, R. A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J. E., Janson, C. A., McElroy, H. E., Gribskov, C. L. & Worland, S. (1994). *Structure of human rhinovirus 3C protease reveals a trypsin-like polypeptide fold, RNA-binding site, and means for cleaving precursor polyprotein. Cell*, **77**, 761–771.

Meng, W., Sawasdikosol, S., Burakoff, S. J. & Eck, M. J. (1999). *Structure of the amino-terminal domain of Cbl complexed to its binding site on ZAP-70 kinase. Nature (London)*, **398**, 84–90.

Merritt, E. A., Sarfaty, S., van den Akker, F., L'hoir, C., Martial, J. A. & Hol, W. G. J. (1994). *Crystal structure of cholera toxin B-pentamer bound to receptor* G_{M1} *pentasaccharide. Protein Sci.* **3**, 166–175.

Merritt, E. A., Sarfaty, S., Feil, I. K. & Hol, W. G. J. (1997). *Structural foundation for the design of receptor antagonists targeting E. coli heat-labile enterotoxin. Structure*, **5**, 1485–1499.

Metcalf, P. & Fusek, M. (1993). *Two crystal structures for cathepsin D: the lysosomal targeting signal and active site. EMBO J.* **12**, 1293–1302.

Mikami, B., Adachi, M., Kage, T., Sarikaya, E., Nanmori, T., Shinke, R. & Utsumi, S. (1999). *Structure of raw starch-digesting Bacillus cereus beta-amylase complexed with maltose. Biochemistry*, **38**, 7050–7061.

Mikol, V., Ma, D. & Carlow, C. K. (1998). *Crystal structure of the cyclophilin-like domain from the parasitic nematode Brugia malayi. Protein Sci.* **7**, 1310–1316.

Milburn, M. V., Hassell, A. M., Lambert, M. H., Jordan, S. R., Proudfoot, A. E. I., Graber, P. & Wells, T. N. C. (1993). *A novel dimer configuration revealed by the crystal structure at 2.4-Å resolution of human interleukin-5. Nature (London)*, **363**, 172–176.

Miller, M. D., Tanner, J., Alpaugh, M., Benedik, M. J. & Krause, K. L. (1994). *2.1 Å structure of Serratia endonuclease suggests a mechanism for binding to double-stranded DNA. Nature Struct. Biol.* **1**, 461–468.

Minke, W. E., Hong, F., Verlinde, C. L. M. J., Hol, W. G. J. & Fan, E. (1999). *Using a galactose library for exploration of a novel hydrophobic pocket in the receptor binding site of the E. coli heat-labile enterotoxin. J. Biol. Chem.* **274**, 33469–33473.

Miyatake, H., Hata, Y., Fujii, T., Hamada, K., Morihara, K. & Katsube, Y. (1995). *Crystal structure of the unliganded alkaline protease from Pseudomonas aeruginosa IFO3080 and its conformational changes on ligand binding. J. Biochem. (Tokyo)*, **118**, 474–479.

Moser, J., Gerstel, B., Meyer, J. E., Chakraborty, T., Wehland, J. & Heinz, D. W. (1997). *Crystal structure of the phosphatidylinositol-specific phospholipase C from the human pathogen Listeria monocytogenes. J. Mol. Biol.* **273**, 269–282.

Mottonen, J., Strand, A., Symersky, J., Sweet, R. M., Danley, D. E., Gioghegan, K. F., Gerard, R. D. & Goldsmith, E. J. (1992). *Structural basis of latency in plasminogen activator inhibitor-1. Nature (London)*, **355**, 270–273.

Muchmore, S. W., Sattler, M., Liang, H., Meadows, R. P., Harlan, J. E., Yoon, H. S., Nettesheim, D., Chang, B. S., Thompson, C. B., Wong, S. L., Ng, S. L. & Fesik, S. W. (1996). *X-ray and NMR structure of human Bcl-xL, an inhibitor of programmed cell death. Nature (London)*, **381**, 335–341.

Mulichak, A. M., Tulinsky, A. & Ravichandran, K. G. (1991). *Crystal and molecular structure of human plasminogen kringle 4 refined at 1.9-Å resolution. Biochemistry*, **30**, 10576–10588.

Mulichak, A. M., Wilson, J. E., Padmanabhan, K. & Garavito, R. M. (1998). *The structure of mammalian hexokinase-1. Nature Struct. Biol.* **5**, 555–560.

Muller, Y. A., Ultsch, M. H. & DeVos, A. M. (1996). *The crystal structure of the extracellular domain of human tissue factor refined to 1.7-Å resolution. J. Mol. Biol.* **256**, 144–159.

Muller, Y. A., Ultsch, M. H., Kelley, R. F. & DeVos, A. M. (1994). *Structure of the extracellular domain of human tissue factor: location of the factor VIIa binding site. Biochemistry*, **33**, 10864–10870.

Murray, C. J. & Salomon, J. A. (1998). *Modeling the impact of global tuberculosis control strategies. Proc. Natl Acad. Sci. USA*, **95**, 13881–13886.

Murray, I. A., Cann, P. A., Day, P. J., Derrick, J. P., Sutcliffe, M. J., Shaw, W. V. & Leslie, A. G. (1995). *Steroid recognition by chloramphenicol acetyltransferase: engineering and structural analysis of a high affinity fusidic acid binding site. J. Mol. Biol.* **254**, 993–1005.

Murray, M. G., Kuhn, R. J., Arita, M., Kawamura, N., Nomoto, A. & Wimmer, E. (1988). *Poliovirus type 1/type 3 antigenic hybrid virus constructed in vitro elicits type 1 and type3 neutralizing antibodies in rabbits and monkeys. Proc. Natl Acad. Sci. USA*, **85**, 3203–3207.

Murthy, H. M., Clum, S. & Padmanabhan, R. (1999). *Dengue virus NS3 serine protease. Crystal structure and insights into interaction of the active site with substrates by molecular modeling and structural analysis of mutational effects. J. Biol. Chem.* **274**, 5573–5580.

Musil, D., Zucic, D., Turk, D., Engh, R. A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katunuma, N. & Bode, W. (1991). *The refined 2.15 Å X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. EMBO J.* **10**, 2321–2330.

Nagar, B., Jones, R. G., Diefenbach, R. J., Isenman, D. E. & Rini, J. M. (1998). *X-ray crystal structure of C3d: a C3 fragment and ligand for complement receptor 2. Science*, **280**, 1277–1281.

Nam, H. J., Haser, W. G., Roberts, T. M. & Frederick, C. A. (1996). *Intramolecular interactions of the regulatory domains of the*

Bcr-Abl kinase reveal a novel control mechanism. Structure, **4**, 1105–1114.

Narayana, N., Matthews, D. A., Howell, E. E. & Nguyen-huu, X. (1995). *A plasmid-encoded dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D2-symmetric active site. Nature Struct. Biol.* **2**, 1018–1025.

National Institute of Allergy and Infectious Diseases (1998). *The Jordan Report: Accelerated Development of Vaccines.* NIAID, Bethesda, MD.

Navia, M. A., Fitzgerald, P. M., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L. & Springer, J. P. (1989). *Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. Nature (London)*, **337**, 615–620.

Navia, M. A., McKeever, B. M., Springer, J. P., Lin, T. Y., Williams, H. R., Fluder, E. M., Dorn, C. P. & Hoogsteen, K. (1989). *Structure of human neutrophil elastase in complex with a peptide chloromethyl ketone inhibitor at 1.84-Å resolution. Proc. Natl Acad. Sci. USA*, **86**, 7–11.

Naylor, C. E., Eaton, J. T., Howells, A., Justin, N., Moss, D. S., Titball, R. W. & Basak, A. K. (1998). *Structure of the key toxin in gas gangrene. Nature Struct. Biol.* **5**, 738–746.

Nurizzo, D., Silvestrini, M. C., Mathieu, M., Cutruzzola, F., Bourgeois, D., Fulop, V., Hajdu, J., Brunori, M., Tegoni, M. & Cambillau, C. (1997). *N-terminal arm exchange is observed in the 2.15 Å crystal structure of oxidized nitrite reductase from Pseudomonas aeruginosa. Structure*, **5**, 1157–1171.

Oefner, C., D'Arcy, A. & Winkler, F. K. (1988). *Crystal structure of human dihydrofolate reductase complexed with folate. Eur. J. Biochem.* **174**, 377–385.

Oinonen, C., Tikkanen, R., Rouvinen, J. & Peltonen, L. (1995). *Three-dimensional structure of human lysosomal aspartylglucosaminidase. Nature Struct. Biol.* **2**, 1102–1108.

Ozaki, H., Sato, T., Kubota, H., Hata, Y., Katsube, Y. & Shimonishi, Y. (1991). *Molecular structure of the toxin domain of heat-stable enterotoxin produced by a pathogenic strain of Escherichia coli. A putative binding site for a binding protein on rat intestinal epithelial cell membranes. J. Biol. Chem.* **266**, 5934–5941.

Padmanabhan, K., Padmanabhan, K. P., Tulinsky, A., Park, C. H., Bode, W., Huber, R., Blankenship, D. T., Cardin, A. D. & Kisiel, W. (1993). *Structure of human des(1–45) factor Xa at 2.2-Å resolution. J. Mol. Biol.* **232**, 947–966.

Pai, E. F., Kabsch, W., Krengel, U., Holmes, K. C., John, J. & Wittinghofer, A. (1989). *Structure of the guanine-nucleotide-binding domain of the Ha-ras oncogene product p21 in the triphosphate conformation. Nature (London)*, **341**, 209–214.

Papageorgiou, A. C., Acharya, K. R., Shapiro, R., Passalacqua, E. F., Brehm, R. D. & Tranter, H. S. (1995). *Crystal structure of the superantigen enterotoxin C2 from Staphylococcus aureus reveals a zinc-binding site. Structure*, **3**, 769–779.

Papageorgiou, A. C., Tranter, H. S. & Acharya, K. R. (1998). *Crystal structure of microbial superantigen staphylococcal enterotoxin B at 1.5 Å resolution: implications for superantigen recognition by MHC class II molecules and T-cell receptors. J. Mol. Biol.* **277**, 61–79.

Pares, S., Mouz, N., Petillot, Y., Hakenbeck, R. & Dideberg, O. (1996). *X-ray structure of Streptococcus pneumoniae PBP2x, a primary penicillin target enzyme. Nature Struct. Biol.* **3**, 284–289.

Parge, H. E., Forest, K. T., Hickey, M. J., Christensen, D. A., Getzoff, E. D. & Tainer, J. A. (1995). *Structure of the fibre-forming protein pilin at 2.6 Å resolution. Nature (London)*, **378**, 32–38.

Parge, H. E., Hallewell, R. A. & Tainer, J. A. (1992). *Atomic structures of wild-type and thermostable mutant recombinant human Cu,Zn superoxide dismutase. Proc. Natl Acad. Sci. USA*, **89**, 6109–6113.

Patskovsky, Y. V., Patskovska, L. N. & Listowsky, I. (1999). *Functions of His107 in the catalytic mechanism of human glutathione s-transferase hGSTM1a-1a. Biochemistry*, **38**, 1193–1202.

Pauptit, R. A., Karlsson, R., Picot, D., Jenkins, J. A., Niklaus-Reimer, A. S. & Jansonius, J. N. (1988). *Crystal structure of neutral protease from Bacillus cereus refined at 3.0 Å resolution and comparison with the homologous but more thermostable enzyme thermolysin. J. Mol. Biol.* **199**, 525–537.

Pearl, L., O'Hara, B., Drew, R. & Wilson, S. (1994). *Crystal structure of AmiC: the controller of transcription antitermination in the amidase operon of Pseudomonas aeruginosa. EMBO J.* **13**, 5810–5817.

Pedelacq, J. D., Maveyraud, L., Prevost, G., Bab-Moussa, L., Gonzalez, A., Coucelle, E., Shepard, W., Monteil, H., Samama, J. P. & Mourey, L. (1999). *The structure of a Staphylococcus aureus leucocidin component (LukF-PV) reveals the fold of the water-soluble species of a family of transmembrane pore-forming toxins. Struct. Fold. Des.* **7**, 277–287.

Pedersen, L. C., Benning, M. M. & Holden, H. M. (1995). *Structural investigation of the antibiotic and ATP-binding sites in kanamycin nucleotidyltransferase. Biochemistry*, **34**, 13305–13311.

Perrakis, A., Tews, I., Dauter, Z., Oppenheim, A. B., Chet, I., Wilson, K. S. & Vorgias, C. E. (1994). *Crystal structure of a bacterial chitinase at 2.3 Å resolution. Structure*, **2**, 1169–1180.

Perutz, M. (1992). *Protein Structure. New Approaches to Disease and Therapy.* New York: W. H. Freeman & Co.

Petosa, C., Collier, R. J., Klimpel, K. R., Leppla, S. H. & Liddington, R. C. (1997). *Crystal structure of the anthrax toxin protective antigen. Nature (London)*, **385**, 833–838.

Pfuegl, G., Kallen, J., Schirmer, T., Jansonius, J. N., Zurini, M. G. M. & Walkinshaw, M. D. (1993). *X-ray structure of a decameric cyclophilin–cyclosporin crystal complex. Nature (London)*, **361**, 91–94.

Phillips, C., Dohnalek, J., Gover, S., Barrett, M. P. & Adams, M. J. (1998). *A 2.8 Å resolution structure of 6-phosphogluconate dehydrogenase from the protozoan parasite Trypanosoma brucei: comparison with the sheep enzyme accounts for differences in activity with coenzyme and substrate analogues. J. Mol. Biol.* **282**, 667–681.

Phillips, C. L., Ullman, B., Brennan, R. G. & Hill, C. P. (1999). *Crystal structures of adenine phosphoribosyltransferase from Leishmania donovani. EMBO J.* **18**, 3533–3545.

Pineo, G. F. & Hull, R. D. (1999). *Thrombin inhibitors as anticoagulant agents. Curr. Opin. Hematol.* **6**, 298–303.

Poljak, R. J., Amzel, L. M., Avey, H. P., Chen, B. L., Phizackerley, R. P. & Saul, F. (1973). *Three-dimensional structure of the Fab′ fragment of a human immunoglobulin at 2.8-Å resolution. Proc. Natl Acad. Sci. USA*, **70**, 3305–3310.

Porter, R. (1999). *The Greatest Benefit to Mankind: a Medical History of Humanity.* New York: W. W. Norton and Co., Inc.

Prasad, G. S., Earhart, C. A., Murray, D. L., Novick, R. P., Schlievert, P. M. & Ohlendorf, D. H. (1993). *Structure of toxic shock syndrome toxin 1. Biochemistry*, **32**, 13761–13766.

Pratt, K. P., Cote, H. C., Chung, D. W., Stenkamp, R. E. & Davie, E. W. (1997). *The primary fibrin polymerization pocket: three-dimensional structure of a 30-kDa C-terminal gamma chain fragment complexed with the peptide Gly-Pro-Arg-Pro. Proc. Natl Acad. Sci. USA*, **94**, 7176–7181.

Pratt, K. P., Shen, B. W., Takeshima, K., Davie, E. W., Fujikawa, K. & Stoddard, B. L. (1999). *Structure of the C2 domain of human factor VIII at 1.5 Å resolution. Nature (London)*, **402**, 439–442.

Priestle, J. P., Schar, H. P. & Grutter, M. G. (1988). *Crystal structure of the cytokine interleukin-1 beta. EMBO J.* **7**, 339–343.

Qiu, X., Culp, J. S., DiLella, A. G., Hellmig, B., Hoog, S. S., Janson, C. A., Smith, W. W. & Abdel-Meguid, S. S. (1996). *Unique fold and active site in cytomegalovirus protease. Nature (London)*, **383**, 275–279.

Qiu, X., Janson, C. A., Culp, J. S., Richardson, S. B., Debouck, C., Smith, W. W. & Abdel-Meguid, S. S. (1997). *Crystal structure of varicella-zoster virus protease. Proc. Natl Acad. Sci. USA*, **94**, 2874–2879.

Qiu, X., Verlinde, C. L. M. J., Zhang, S., Schmitt, M. P., Holmes, R. K. & Hol, W. G. J. (1995). *Three-dimensional structure of the diphtheria toxin repressor in complex with divalent cation co-repressors. Structure*, **3**, 87–100.

Rabijns, A., De Bondt, H. L. & De Ranter, C. (1997). *Three-dimensional structure of staphylokinase, a plasminogen activator with therapeutic potential. Nature Struct. Biol.* **4**, 357–360.

Radhakrishnan, R., Walter, L. J., Hruza, A., Reichert, P., Trotta, P. P., Nagabhushan, T. L. & Walter, M. R. (1996). *Zinc mediated dimer of human interferon-alpha 2b revealed by X-ray crystallography. Structure*, **4**, 1453–1463.

Raghunathan, S., Chandross, R. J., Kretsinger, R. H., Allison, T. J., Penington, C. J. & Rule, G. S. (1994). *Crystal structure of human class mu glutathione transferase GSTM2–2. Effects of lattice packing on conformational heterogeneity. J. Mol. Biol.* **238**, 815–832.

Rano, T. A., Timkey, T., Peterson, E. P., Rotonda, J., Nicholson, D. W., Becker, J. W., Chapman, K. T. & Thornberry, N. A. (1997). *A combinatorial approach for determining protease specificities: application to interleukin-1 beta converting enzyme (ICE). Chem. Biol.* **4**, 149–155.

Rao, Z., Handford, P., Mayhew, M., Knott, V., Brownlee, G. G. & Stuart, D. (1995). *The structure of a Ca(2+)-binding epidermal growth factor-like domain: its role in protein–protein interactions. Cell*, **82**, 131–141.

Read, J. A., Wilkinson, K. W., Tranter, R., Sessions, R. B. & Brady, R. L. (1999). *Chloroquine binds in the cofactor binding site of Plasmodium falciparum lactate dehydrogenase. J. Biol. Chem.* **274**, 10213–10218.

Redinbo, M. R., Stewart, L., Kuhn, P., Champoux, J. J. & Hol, W. G. J. (1998). *Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science*, **279**, 1504–1513.

Reichmann, L., Clark, M., Waldmann, H. & Winter, G. (1988). *Reshaping human antibodies for therapy. Nature (London)*, **332**, 323–327.

Reinemer, P., Dirr, H. W., Ladenstein, R., Huber, R., Lo Bello, M., Federici, G. & Parker, M. W. (1992). *Three-dimensional structure of class pi glutathione S-transferase from human placenta in complex with S-hexylglutathione at 2.8 Å resolution. J. Mol. Biol.* **227**, 214–226.

Reinemer, P., Grams, F., Huber, R., Kleine, T., Schnierer, S., Piper, M., Tschesche, H. & Bode, W. (1994). *Structural implications for the role of the N terminus in the 'superactivation' of collagenases. A crystallographic study. FEBS Lett.* **338**, 227–233.

Ren, J., Esnouf, R., Garman, E., Somers, D., Ross, C., Kirby, I., Keeling, J., Darby, G., Jones, Y., Stuart, D. & Stammers, D. (1995). *High resolution structures of HIV-1 RT from four RT-inhibitor complexes. Nature Struct. Biol.* **2**, 293–302.

Ren, J., Esnouf, R. M., Hopkins, A. L., Jones, E. Y., Kirby, I., Keeling, J., Ross, C. K., Larder, B. A., Stuart, D. I. & Stammers, D. K. (1998). *3-Azido-3′-deoxythymidine drug resistance mutations in HIV-1 reverse transcriptase can induce long range conformational changes. Proc. Natl Acad. Sci. USA*, **95**, 9518–9523.

Renwick, S. B., Snell, K. & Baumann, U. (1998). *The crystal structure of human cytosolic serine hydroxymethyltransferase: a target for cancer chemotherapy. Structure*, **15**, 1105–1116.

Resnick, D. A., Smith, A. D., Geisler, S. C., Zhang, A., Arnold, E. & Arnold, G. F. (1995). *Chimeras from a human rhinovirus 14–human immunodeficiency virus type 1 (HIV-1) V3 loop seroprevalence library induce neutralizing responses against HIV-1. J. Virol.* **69**, 2406–2411.

Rey, F. A., Heinz, F. X., Mandl, C., Kunz, C. & Harrison, S. C. (1995). *The envelope glycoprotein from tick-borne encephalitis virus at 2 Å resolution. Nature (London)*, **375**, 291–298.

Rigden, D. J., Phillips, S. E., Michels, P. A. & Fothergill-Gilmore, L. A. (1999). *The structure of pyruvate kinase from Leishmania mexicana reveals details of the allosteric transition and unusual effector specificity. J. Mol. Biol.* **291**, 615–635.

Ripka, W. C. (1997). *Design of antithrombotic agents directed at factor Xa*. In *Structure-Based Drug Design*, edited by P. Veerapandian, pp. 265–294. New York: Marcel Dekker.

Roberts, M. M., White, J. L., Grutter, M. G. & Burnett, R. M. (1986). *Three-dimensional structure of the adenovirus major coat protein hexon. Science*, **232**, 1148–1151.

Rodgers, D. W., Bamblin, S. J., Harris, B. A., Ray, S., Culp, J. S., Hellmig, B., Woolf, D. J., Debouck, C. & Harrison, S. C. (1995). *The structure of unliganded reverse transcriptase from the human immunodeficiency virus type 1. Proc. Natl Acad. Sci. USA*, **92**, 1222–1226.

Roe, S. M., Barlow, T., Brown, T., Oram, M., Keeley, A., Tsaneva, I. R. & Pearl, L. H. (1998). *Crystal structure of an octameric RuvA–Holliday junction complex. Mol. Cell*, **2**, 361–372.

Rolan, P. E., Parker, J. E., Gray, S. J., Weatherley, B. C., Ingram, J., Leavens, W., Wootton, R. & Posner, J. (1993). *The pharmacokinetics, tolerability and pharmacodynamics of tucaresol (589C80; 4[2-formyl-3-hydroxyphenoxymethyl] benzoic acid), a potential anti-sickling agent, following oral administration to healthy subjects. Br. J. Clin. Pharmacol.* **35**, 419–425.

Rossjohn, J., Feil, S. C., McKinstry, W. J., Tweten, R. K. & Parker, M. W. (1997). *Structure of a cholesterol-binding, thiol-activated cytolysin and a model of its membrane form. Cell*, **89**, 685–692.

Rossjohn, J., Feil, S. C., Wilce, M. C. J., Sexton, J. L., Spithill, T. W. & Parker, M. W. (1997). *Crystallization, structural determination and analysis of a novel parasite vaccine candidate: Fasciola hepatica glutathione S-transferase. J. Mol. Biol.* **273**, 857–872.

Rossjohn, J., McKinstry, W. J., Oakley, A. J., Verger, D., Flanagan, J., Chelvanayagam, G., Tan, K. L., Board, P. G. & Parker, M. W. (1998). *Human theta class glutathione transferase: the crystal structure reveals a sulfate-binding pocket within a buried active site. Structure*, **6**, 309–322.

Rossjohn, J., Polekhina, G., Feil, S. C., Allocati, N., Masulli, M., De Illio, C. & Parker, M. W. (1998). *A mixed disulfide bond in bacterial glutathione transferase: functional and evolutionary implications. Structure*, **6**, 721–734.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Structure of a human common cold virus and functional relationship to other picornaviruses. Nature (London)*, **317**, 145–153.

Roussel, A., Anderson, B. F., Baker, H. M., Fraser, J. D. & Baker, E. N. (1997). *Crystal structure of the streptococcal superantigen SPE-C: dimerization and zinc binding suggest a novel mode of interaction with MHC class II molecules. Nature Struct. Biol.* **4**, 635–643.

Rudenko, G., Bonten, E., d'Azzo, A. & Hol, W. G. J. (1995). *Three-dimensional structure of the human 'protective protein': structure of the precursor form suggests a complex activation mechanism. Structure*, **3**, 1249–1259.

Rudenko, G., Bonten, E., Hol, W. G. J. & d'Azzo, A. (1998). *The atomic model of the human protective protein/cathepsin A suggests a structural basis for galactosialidosis. Proc. Natl Acad. Sci. USA*, **95**, 621–625.

Russo, A. A., Tong, L., Lee, J. O., Jeffrey, P. D. & Pavletich, N. P. (1998). *Structural basis for inhibition of the cyclin-dependent kinase Cdk6 by the tumour suppressor p16INK4a. Nature (London)*, **395**, 237–243.

Rydel, T. J., Ravichandran, K. G., Tulinsky, A., Bode, W., Huber, R., Roitsch, C. & Fenton, J. W. I. (1990). *The structure of a complex of recombinant hirudin and human alpha-thrombin. Science*, **249**, 277–280.

Rydel, T. J., Yin, M., Padmanabhan, K. P., Blankenship, D. T., Cardin, A. D., Correa, P. E., Fenton, J. W. I. & Tulinsky, A. (1994). *Crystallographic structure of human gamma-thrombin. J. Biol. Chem.* **269**, 22000–22006.

Sarafianos, S. G., Das, K., Ding, J., Boyer, P. L., Hughes, S. H. & Arnold, E. (1999). *Touching the heart of HIV-1 drug resistance: the fingers close down on the dNTP at the polymerase active site. Chem. Biol.* **6**, R137–R146.

Sauer, F. G., Futterer, K., Pinkner, J. S., Dodson, K. W., Hultgren, S. J. & Waksman, G. (1999). *Structural basis of chaperone function and pilus biogenesis. Science*, **285**, 1058–1061.

Savva, R., McAuley-Hecht, K., Brown, T. & Pearl, L. (1995). *The structural basis of specific base-excision repair by uracil-DNA glycosylase. Nature (London)*, **373**, 487–493.

Scheffzek, K., Ahmadian, M. R., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F. & Wittinghofer, A. (1997). *The Ras–RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. Science*, **277**, 333–338.

Schiffer, C. A., Clifton, I. J., Davisson, V. J., Santi, D. V. & Stroud, R. M. (1995). *Crystal structure of human thymidylate synthase: a structural mechanism for guiding substrates into the active site. Biochemistry*, **34**, 16279–16287.

Schlagenhauf, E., Etges, R. & Metcalf, P. (1998). *The crystal structure of the Leishmania major surface proteinase leishmanolysin (gp63). Structure*, **6**, 1035–1046.

Schonbrunn, E., Sack, S., Eschenberg, S., Perrakis, A., Krekel, F., Amrhein, N. & Mandelkow, E. (1996). *Crystal structure of UDP-N-acetylglucosamine enolpyruvyltransferase, the target of the antibiotic fosfomycin. Structure*, **4**, 1065–1075.

Schreuder, H., Tardif, C., Trump-Kallmeyer, S., Soffientini, A., Sarubbi, E., Akeson, A., Bowlin, T., Yanofsky, S. & Barrett, R. W. (1997). *A new cytokine-receptor binding mode revealed by the crystal structure of the IL-1 receptor with an antagonist. Nature (London)*, **386**, 194–200.

Schreuder, H. A., de Boer, B., Dijkema, R., Mulders, J., Theunissen, H. J. M., Grootenhuis, P. D. J. & Hol, W. G. J. (1994). *The intact and cleaved human antithrombin III complex as a model for serpin-proteinase interactions. Nature Struct. Biol.* **1**, 48–54.

Schumacher, M. A., Carter, D., Ross, D. S., Ullman, B. & Brennan, R. G. (1996). *Crystal structures of Toxoplasma gondii HGXPRTase reveal the catalytic role of a long flexible loop. Nature Struct. Biol.* **3**, 881–887.

Schumacher, M. A., Carter, D., Scott, D. M., Roos, D. S., Ullman, B. & Brennan, R. G. (1998). *Crystal structures of Toxoplasma gondii uracil phosphoribosyltransferase reveal the atomic basis of pyrimidine discrimination and prodrug binding. EMBO J.* **17**, 3219–3232.

Schwabe, J. W. E., Chapman, L., Finch, J. T. & Rhodes, D. (1993). *The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. Cell*, **75**, 567–578.

Scott, D. L., White, S. P., Browning, J. L., Rosa, J. J., Gelb, M. H. & Sigler, P. B. (1991). *Structures of free and inhibited human secretory phospholipase A2 from inflammatory exudate. Science*, **254**, 1007–1010.

Sha, B. & Luo, M. (1997). *Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. Nature Struct. Biol.* **4**, 239–244.

Shah, S. A., Shen, B. W. & Brunger, A. T. (1997). *Human ornithine aminotransferase complexed with L-canaline and gabaculine: structural basis for substrate recognition. Structure*, **5**, 1067–1075.

Sharma, A., Hanai, R. & Mondragon, A. (1994). *Crystal structure of the amino-terminal fragment of vaccinia virus DNA topoisomerase I at 1.6 Å resolution. Structure*, **2**, 767–777.

Sharma, V., Grubmeyer, C. & Sacchettini, J. C. (1998). *Crystal structure of quinolinic acid phosphoribosyltransferase from Mycobacterium tuberculosis: a potential TB drug target. Structure*, **6**, 1587–1599.

Sherry, B., Mosser, A. G., Colonno, R. J. & Rueckert, R. R. (1986). *Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. J. Virol.* **57**, 246–257.

Sherry, B. & Rueckert, R. (1985). *Evidence for at least two dominant neutralization antigens on human rhinovirus 14. J. Virol.* **53**, 137–143.

Shi, D., Morizono, H., Ha, Y., Aoyagi, M., Tuchman, M. & Allewell, N. M. (1998). *1.85-Å resolution crystal structure of human ornithine transcarbamoylase complexed with N-phosphonacetyl-L-ornithine. Catalytic mechanism and correlation with inherited deficiency. J. Biol. Chem.* **273**, 34247–34254.

Shi, W., Li, C. M., Tyler, P. C., Furneaux, R. H., Cahill, S. M., Girvin, M. E., Grubmeyer, C., Schramm, V. L. & Almo, S. C. (1999). *The 2.0 Å structure of malarial purine phosphoribosyltransferase in complex with a transition-state analogue inhibitor. Biochemistry*, **38**, 9872–9880.

Shi, W., Schramm, V. L. & Almo, S. C. (1999). *Nucleoside hydrolase from Leishmania major. Cloning, expression, catalytic properties, transition state inhibitors, and the 2.5-Å crystal structure. J. Biol. Chem.* **274**, 21114–21120.

Shieh, H. S., Kurumbail, R. G., Stevens, A. M., Stegeman, R. A., Sturman, E. J., Pak, J. Y., Wittwer, A. J., Palmier, M. O., Wiegand, R. C., Holwerda, B. C. & Stallings, W. C. (1996). *Three-dimensional structure of human cytomegalovirus protease. Nature (London)*, **383**, 279–282.

Sielecki, A. R., Hayakawa, K., Fujinaga, M., Murphy, M. E. P., Fraser, M., Muir, A. K., Carilli, C. T., Lewicki, J. A., Baxter, J. D. & James, M. N. G. (1989). *Structure of recombinant human renin, a target for cardiovascular-active drugs, at 2.5-Å resolution. Science*, **243**, 1346–1351.

Silva, A. M., Lee, A. Y., Gulnik, S. V., Maier, P., Collins, J., Bhat, T. N., Collins, P. J., Cachau, R. E., Luker, K. E., Gluzman, I. Y., Francis, S. E., Oksman, A., Goldberg, D. E. & Erickson, J. W. (1996). *Structure and inhibition of plasmepsin II, a hemoglobin-degrading enzyme from Plasmodium falciparum. Proc. Natl Acad. Sci. USA*, **93**, 10034–10039.

Silvian, L. F., Wang, J. & Steitz, T. A. (1999). *Insights into editing from an ile-tRNA synthetase structure with tRNAile and mupirocin. Science*, **285**, 1074–1077.

Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., Board, P. G., Olin, B., Mannervik, B. & Jones, T. A. (1993). *Structure determination and refinement of human alpha class glutathione transferase A1-1, and a comparison with the mu and pi class enzymes. J. Mol. Biol.* **232**, 192–212.

Sixma, T. K., Pronk, S. E., Kalk, K. H., Wartna, E. S., van Zanten, B. A. M., Witholt, B. & Hol, W. G. J. (1991). *Crystal structure of a cholera toxin-related heat-labile enterotoxin from E. coli. Nature (London)*, **351**, 371–377.

Skinner, R., Abrahams, J. P., Whisstock, J. C., Lesk, A. M., Carrell, R. W. & Wardell, M. R. (1997). *The 2.6 Å structure of antithrombin indicates a conformational change at the heparin binding site. J. Mol. Biol.* **266**, 601–609.

Skinner, R., Chang, W. S. W., Jin, L., Pei, X. Y., Huntington, J. A., Abrahams, J. P., Carrell, R. W. & Lomas, D. A. (1998). *Implications for function and therapy of a 2.9 Å structure of binary-complexed antithrombin. J. Mol. Biol.* **283**, 9–14.

Smith, A. D., Geisler, S. C., Chen, A. A., Resnick, D. A., Roy, B. M., Lewi, P. J., Arnold, E. & Arnold, G. F. (1998). *Human rhinovirus type 14:human immunodeficiency virus type 1 (HIV-1) V3 loop chimeras from a combinatorial library induce potent neutralizing antibody responses against HIV-1. J. Virol.* **72**, 651–659.

Somers, W., Ultsch, M., DeVos, A. M. & Kossiakoff, A. A. (1994). *The X-ray structure of a growth hormone–prolactin receptor complex. Nature (London)*, **372**, 478–481.

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. & Gouaux, J. E. (1996). *Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science*, **274**, 1859–1866.

Souza, D. H., Garratt, R. C., Araujo, A. P., Guimaraes, B. G., Jesus, W. D., Michels, P. A., Hannaert, V. & Oliva, G. (1998). *Trypanosoma cruzi glycosomal glyceraldehyde-3-phosphate dehydrogenase: structure, catalytic mechanism and targeted inhibitor design. FEBS Lett.* **424**, 131–135.

Spraggon, G., Everse, S. J. & Doolittle, R. F. (1997). *Crystal structures of fragment D from human fibrinogen and its crosslinked counterpart from fibrin. Nature (London)*, **389**, 455–462.

Spraggon, G., Phillips, C., Nowak, U. K., Ponting, C. P., Saunders, D., Dobson, C. M., Stuart, D. I. & Jones, E. Y. (1995). *The crystal structure of the catalytic domain of human urokinase-type plasminogen activator. Structure*, **3**, 681–691.

Spurlino, J. C., Smallwood, A. M., Carlton, D. D., Banks, T. M., Vavra, K. J., Johnson, J. S., Cook, E. R., Falvo, J., Wahl, R. D., Pulvino, T. A., Wendoloski, J. J. & Smith, D. L. (1994). *1.56-Å structure of mature truncated human fibroblast collagenase. Proteins*, **19**, 98–109.

Stams, T., Spurlino, J. C., Smith, D. L., Wahl, R. C., Ho, T. F., Qoronfleh, M. H., Banks, T. M. & Rubin, B. (1994). *Structure of human neutrophil collagenase reveals large S1′ specificity pocket. Nature Struct. Biol.* **1**, 119–123.

Stein, P. E., Boodhoo, A., Armstrong, G. D., Cockle, S. A., Klein, M. H. & Read, R. J. (1994). *The crystal structure of pertussis toxin. Structure*, **2**, 45–57.

Stein, P. E., Boodhoo, A., Tyrrell, G. J., Brunton, J. L. & Read, R. J. (1992). *Crystal structure of the cell-binding B oligomer of verotoxin-1 from E. coli. Nature (London)*, **355**, 748–750.

Stewart, L., Redinbo, M. R., Qiu, X., Hol, W. G. J. & Champoux, J. J. (1998). *A model for the mechanism of human topoisomerase I. Science*, **279**, 1534–1541.

Stubbs, M. T., Laber, B., Bode, W., Huber, R., Jerala, R., Lenarcic, B. & Turk, V. (1990). *The refined 2.4 Å X-ray crystal structure of recombinant human stefin B in complex with the cysteine proteinase papain: a novel type of proteinase inhibitor interaction. EMBO J.* **9**, 1939–1947.

Stuckey, J. A., Schubert, H. L., Fauman, E. B., Zhang, Z. Y., Dixon, J. E. & Saper, M. A. (1994). *Crystal structure of Yersinia protein tyrosine phosphatase at 2.5 Å and the complex with tungstate. Nature (London)*, **370**, 571–575.

Sugio, S., Kashima, A., Mochizuki, S., Noda, M. & Kobayashi, K. (1999). *Crystal structure of human serum albumin at 2.5 Å resolution. Protein Eng.* **12**, 439–446.

Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987). *Structure and refinement at 1.8 Å resolution of the aspartic proteinase from Rhizopus chinensis. J. Mol. Biol.* **196**, 877–900.

Sundstrom, M., Hallen, D., Svensson, A., Schad, E., Dohlsten, M. & Abrahmsen, L. (1996). *The co-crystal structure of staphylococcal enterotoxin type A with Zn^{2+} at 2.7 Å resolution. Implications for major histocompatibility complex class II binding. J. Biol. Chem.* **271**, 32212–32216.

Symersky, J., Patti, J. M., Carson, M., House-Pompeo, K., Teale, M., Moore, D., Jin, L., Schneider, A., DeLucas, L. J., Hook, M. & Narayana, S. V. (1997). *Structure of the collagen-binding domain from a Staphylococcus aureus adhesin. Nature Struct. Biol.* **4**, 833–838.

Taylor, P., Page, A. P., Kontopidis, G., Husi, H. & Walkinshaw, M. D. (1998). *The X-ray structure of a divergent cyclophilin from the nematode parasite Brugia malayi. FEBS Lett.* **425**, 361–366.

Testa, B. (1994). *Drug metabolism.* In *Burger's Medicinal Chemistry and Drug Discovery*, 5th ed., edited by M. E. Wolf, Vol. 1. New York: John Wiley & Sons.

Tews, I., Perrakis, A., Oppenheim, A., Dauter, Z., Wilson, K. S. & Vorgias, C. E. (1996). *Bacterial chitobiase structure provides insight into catalytic mechanism and the basis of Tay–Sachs disease. Nature Struct. Biol.* **3**, 638–648.

Thayer, M. M., Flaherty, K. M. & McKay, D. B. (1991). *Three-dimensional structure of the elastase of Pseudomonas aeruginosa at 1.5-Å resolution. J. Biol. Chem.* **266**, 2864–2871.

Thieme, R., Pai, E. F., Schirmer, R. H. & Schulz, G. E. (1981). *Three-dimensional structure of glutathione reductase at 2 Å resolution. J. Mol. Biol.* **152**, 763–782.

Thompson, S. K., Halbert, S. M., Bossard, M. J., Tomaszek, T. A., Levy, M. A., Zhao, B., Smith, W. W., Abdel-Meguid, S. S., Janson, C. A., D'Alessio, K. J., McQueney, M. S., Amegadzie, B. Y., Hanning, C. R., DesJarlais, R. L., Briand, J., Sarkar, S. K., Huddleston, M. J., James, C. F., Carr, S. A., Garges, K. T., Shu, A., Heys, J. R., Bradbeer, J., Zembryki, D. & Veber, D. F. (1997). *Design of potent and selective* *human cathepsin K inhibitors that span the active site. Proc. Natl Acad. Sci. USA*, **94**, 14249–14254.

Thylefors, B., Negrel, A. D., Pararajasegaram, R. & Dadzie, K. Y. (1995). *Global data on blindness. Bull. WHO*, **73**, 115–121.

Tiffany, K. A., Roberts, D. L., Wang, M., Paschke, R., Mohsen, A. W., Vockley, J. & Kim, J. J. (1997). *Structure of human isovaleryl-CoA dehydrogenase at 2.6 Å resolution: structural basis for substrate specificity. Biochemistry*, **36**, 8455–8464.

Toney, M. D., Hohenester, E., Cowan, S. W. & Jansonius, J. N. (1993). *Dialkylglycine decarboxylase structure: bifunctional active site and alkali metal sites. Science*, **261**, 756–759.

Tong, L., Pav, S., White, D. M., Rogers, S., Crane, K. M., Cywin, C. L., Brown, M. L. & Pargellis, C. A. (1997). *A highly specific inhibitor of human p38 MAP kinase binds in the ATP pocket. Nature Struct. Biol.* **4**, 311–316.

Tong, L., Qian, C., Massariol, M. J., Bonneau, P. R., Cordingley, M. G. & Lagace, L. (1996). *A new serine-protease fold revealed by the crystal structure of human cytomegalovirus protease. Nature (London)*, **383**, 272–275.

Tran, P. H., Korszun, Z. R., Cerritelli, S., Springhorn, S. S. & Lacks, S. A. (1998). *Crystal structure of the DpnM DNA adenine methyltransferase from the Dpnii restriction system of Streptococcus pneumoniae bound to S-adenosylmethionine. Structure*, **6**, 1563–1575.

Tskovsky, Y. V., Patskovska, L. N. & Listowsky, I. (1999). *Functions of His107 in the catalytic mechanism of human glutathione s-transferase hGSTM1a-1a. Biochemistry*, **38**, 1193–1202.

Turner, B. G. & Summers, M. F. (1999). *Structural biology of HIV. J. Mol. Biol.* **285**, 1–32.

Umland, T. C., Wingert, L. M., Swaminathan, S., Furey, W. F., Schmidt, J. J. & Sax, M. (1997). *Structure of the receptor binding fragment HC of tetanus neurotoxin. Nature Struct. Biol.* **4**, 788–792.

Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1991). *Atomic structure of FKBP-FK506, an immunophilin–immunosuppressant complex. Science*, **252**, 839–842.

Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1993). *Atomic structures of the human immunophilin FKBP-12 complexes with FK506 and rapamycin. J. Mol. Biol.* **229**, 105–124.

Van Duyne, G. D., Standaert, R. F., Schreiber, S. L. & Clardy, J. (1991). *Atomic structure of the rapamycin human immunophilin FKBP-12 complex. J. Am. Chem. Soc.* **113**, 7433–7434.

Varghese, J. N., Laver, W. G. & Colman, P. M. (1983). *Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. Nature (London)*, **303**, 35–40.

Varney, M. D., Palmer, C. L., Romines, W. H. R., Boritzki, T., Margosiak, S. A., Almassy, R., Janson, C. A., Bartlett, C., Howland, E. J. & Ferre, R. (1997). *Protein structure-based design, synthesis, and biological evaluation of 5-thia-2,6-diamino-4(3H)-oxopyrimidines: potent inhibitors of glycinamide ribonucleotide transformylase with potent cell growth inhibition. J. Med. Chem.* **40**, 2502–2524.

Vath, G. M., Earhart, C. A., Rago, J. V., Kim, M. H., Bohach, G. A., Schlievert, P. M. & Ohlendorf, D. H. (1997). *The structure of the superantigen exfoliative toxin A suggests a novel regulation as a serine protease. Biochemistry*, **36**, 1559–1566.

Veerapandian, P. (1997). Editor. *Structure-Based Drug Design.* New York: Marcel-Dekker.

Velanker, S. S., Ray, S. S., Gokhale, R. S., Suma, S., Balaram, P. & Murthy, M. R. (1997). *Triosephosphate isomerase from Plasmodium falciparum: the crystal structure provides insights into antimalarial drug design. Structure*, **5**, 751–761.

Vellieux, F. M., Hajdu, J., Verlinde, C. L., Groendijk, H., Read, R. J., Greenhough, T. J., Campbell, J. W., Kalk, K. H., Littlechild, J. A., Watson, H. C. & Hol, W. G. J. (1993). *Structure of glycosomal glyceraldehyde-3-phosphate dehydrogenase from Trypanosoma brucei determined from Laue data. Proc. Natl Acad. Sci. USA*, **90**, 2355–2359.

Verlinde, C. L. M. J. & Hol, W. G. J. (1994). *Structure-based drug design: progress, results and challenges. Structure*, **2**, 577–587.

Vigers, G. P., Anderson, L. J., Caffes, P. & Brandhuber, B. J. (1997). *Crystal structure of the type-I interleukin-1 receptor complexed with interleukin-1beta. Nature (London)*, **386**, 190–194.

Villeret, V., Tricot, C., Stalon, V. & Dideberg, O. (1995). *Crystal structure of Pseudomonas aeruginosa catabolic ornithine transcarbamoylase at 3.0-Å resolution: a different oligomeric organization in the transcarbamoylase family. Proc. Natl Acad. Sci. USA*, **92**, 10762–10766.

Vondrasek, J., van Buskirk, C. P. & Wlodawer, A. (1997). *Database of three-dimensional structure of HIV proteinases. Nature Struct. Biol.* **4**, 8.

Waksman, G., Shoelson, S. E., Pant, N., Cowburn, D. & Kuriyan, J. (1993). *Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. Cell*, **72**, 779–790.

Walker, N. P. C., Talanian, R. V., Brady, K. D., Dang, L. C., Bump, N. J., Ferenz, C. R., Franklin, S., Ghayur, T., Hackett, M. C., Hammill, L. D., Herzog, L., Hugunin, M., Houy, W., Mankovich, J. A., McGuiness, L., Orlewicz, E., Paskind, M., Pratt, C. A., Reis, P., Summani, A., Terranova, M. & Welch, J. P. (1994). *Crystal structure of the cysteine protease interleukin-1 beta-converting enzyme: a (p20/p10)2 homodimer. Cell*, **78**, 343–352.

Walsh, C. T., Fisher, S. L., Park, I. S., Prahalad, M. & Wu, Z. (1996). *Bacterial resistance to vancomycin: five genes and one missing hydrogen bond tell the story. Chem. Biol.* **3**, 21–28.

Walter, M. R., Windsor, W. T., Nagabhushan, T. L., Lundell, D. J., Lunn, C. A., Zauodny, P. J. & Narula, S. K. (1995). *Crystal structure of a complex between interferon-gamma and its soluble high-affinity receptor. Nature (London)*, **376**, 230–235.

Wang, A. H., Ughetto, G., Quigley, G. J. & Rich, A. (1987). *Interactions between an anthracycline antibiotic and DNA: molecular structure of daunomycin complexed to d(CpGpTpApCpG) at 1.2 Å resolution. Biochemistry*, **26**, 1152–1163.

Warner, P., Green, R. C., Gomes, B. & Strimpler, A. M. (1994). *Non-peptidic inhibitors of human leukocyte elastase. 1. The design and synthesis of pyridone-containing inhibitors. J. Med. Chem.* **37**, 3090–3099.

Watanabe, K., Hata, Y., Kizaki, H., Katsube, Y. & Suzuki, Y. (1997). *The refined crystal structure of Bacillus cereus oligo-1,6-glucosidase at 2.0 Å resolution: structural characterization of proline-substitution sites for protein thermostabilization. J. Mol. Biol.* **269**, 142–153.

Weber, P. C. & Czarniecki, M. (1997). *Structure-based design of thrombin inhibitors*. In *Structure-Based Drug Design*, edited by P. Veerapandian, pp. 247–264. New York: Marcel Dekker.

Wei, A. Z., Mayr, I. & Bode, W. (1988). *The refined 2.3-Å crystal structure of human leukocyte elastase in a complex with a valine chloromethyl ketone inhibitor. FEBS Lett.* **234**, 367–373.

Weissenhorn, W., Carfi, A., Lee, K. H., Skehel, J. J. & Wiley, D. C. (1998). *Crystal structure of the Ebola virus membrane fusion subunit, GP2, from the envelope glycoprotein ectodomain. Mol. Cell*, **2**, 605–616.

Wery, J. P., Schevitz, R. W., Clawson, D. K., Bobbitt, J. L., Dow, E. R., Gamboa, G., Goodson, T. J., Hermann, R. B., Kramer, R. M., McClure, D. B., Mihelich, E. D., Putnam, J. E., Sharp, J. D., Stark, D. H., Teater, C., Warrick, M. W. & Jones, N. D. (1991). *Structure of recombinant human rheumatoid arthritic synovial fluid phospholipase A2 at 2.2-Å resolution. Nature (London)*, **352**, 79–82.

Weston, S. A., Camble, R., Colls, J., Rosenbrock, G., Taylor, I., Egerton, M., Tucker, A. D., Tunnicliffe, A., Mistry, A., Macia, F., de La Fortelle, E., Irwin, J., Bricogne, G. & Pauptit, R. A. (1998). *Crystal structure of the anti-fungal target N-myristoyl transferase. Nature Struct. Biol.* **5**, 213–221.

Whitlow, M., Howard, A. J., Stewart, D., Hardman, K. D., Kuyper, L. F., Baccanari, D. P., Fling, M. E. & Tansik, R. L. (1997). *X-ray crystallographic studies of Candida albicans dihydrofolate reductase. High resolution structure of the holoenzyme and an inhibited ternary complex. J. Biol. Chem.* **272**, 30289–30298.

Whittingham, J. L., Edwards, D. J., Antson, A. A., Clarkson, J. M. & Dodson, G. G. (1998). *Interactions of phenol and m-cresol in the insulin hexamer, and their effect on the association properties of B28 pro Asp insulin analogues. Biochemistry*, **37**, 11516–11523.

Whittle, P. J. & Blundell, T. L. (1994). *Protein structure-based drug design. Annu. Rev. Biophys. Biomol. Struct.* **23**, 349–375.

Wierenga, R. K., Kalk, K. H. & Hol, W. G. J. (1987). *Structure determination of the glycosomal triosephosphate isomerase from Trypanosoma brucei brucei at 2.4 Å resolution. J. Mol. Biol.* **198**, 109–121.

Wild, K., Bohner, T., Aubry, A., Folkers, G. & Schulz, G. E. (1995). *The three-dimensional structure of thymidine kinase from herpes simplex virus type 1. FEBS Lett.* **368**, 289–292.

Williams, J. C., Zeelen, J. P., Neubauer, G., Vriend, G., Backmann, J., Michels, P. A., Lambeir, A. M. & Wierenga, R. K. (1999). *Structural and mutagenesis studies of leishmania triosephosphate isomerase: a point mutation can convert a mesophilic enzyme into a superstable enzyme without losing catalytic power. Protein Eng.* **12**, 243–250.

Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F. & McRee, D. E. (2000). *Mammalian microsomal cytochrome P450 monooxygenase:*

structural adaptations for membrane binding and functional diversity. Mol. Cell, **5**, 121–131.

Williams, S. P. & Sigler, P. B. (1998). *Atomic structure of progesterone complexed with its receptor. Nature (London)*, **393**, 392–396.

Wilson, D. K., Bohren, K. M., Gabbay, K. H. & Quiocho, F. A. (1992). *An unlikely sugar substrate site in the 1.65 Å structure of the human aldose reductase holoenzyme implicated in diabetic complications. Science*, **257**, 81–84.

Wilson, I. A., Skehel, J. J. & Wiley, D. C. (1981). *Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature (London)*, **289**, 366–373.

Wilson, K. P., Fitzgibbon, M. J., Caron, P. R., Griffith, J. P., Chen, W., McCaffrey, P. G., Chambers, S. P. & Su, M. S. (1996). *Crystal structure of p38 mitogen-activated protein kinase. J. Biol. Chem.* **271**, 27696–27700.

Wireko, F. C. & Abraham, D. J. (1991). *X-ray diffraction study of the binding of the antisickling agent 12C79 to human hemoglobin. Proc. Natl Acad. Sci. USA*, **88**, 2209–2211.

Wlodawer, A., Miller, M., Jaskolski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. (1989). *Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science*, **245**, 616–621.

Wlodawer, A. & Vondrasek, J. (1998). *Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. Annu. Rev. Biophys. Biomol. Struct.* **27**, 249–284.

Wolf, E., Vassilev, A., Makino, Y., Sali, A., Nakatani, Y. & Burley, S. K. (1998). *Crystal structure of a GCN5-related N-acetyltransferase: Serratia marcescens aminoglycoside 3-N-acetyltransferase. Cell*, **94**, 439–449.

Worthylake, D. K., Wang, H., Yoo, S., Sundquist, W. I. & Hill, C. P. (1999). *Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution. Acta Cryst.* D**55**, 85–92.

Wynne, S. A., Crowther, R. A. & Leslie, A. G. (1999). *The crystal structure of the human hepatitis B virus capsid. Mol. Cell*, **3**, 771–780.

Xia, D., Henry, L. J., Gerard, R. D. & Deisenhofer, J. (1994). *Crystal structure of the receptor-binding domain of adenovirus type 5 fiber protein at 1.7 Å resolution. Structure*, **2**, 1259–1270.

Xie, X., Gu, Y., Fox, T., Coll, J. T., Fleming, M. A., Markland, W., Caron, P. R., Wilson, K. P. & Su, M. S. (1998). *Crystal structure of JNK3: a kinase implicated in neuronal apoptosis. Structure*, **6**, 983–991.

Xu, W., Harrison, S. C. & Eck, M. J. (1997). *Three-dimensional structure of the tyrosine kinase c-Src. Nature (London)*, **385**, 595–602.

Xue, Y., Bjorquist, P., Inghardt, T., Linschoten, M., Musil, D., Sjolin, L. & Deinum, J. (1998). *Interfering with the inhibitory mechanism of serpins: crystal structure of a complex formed between cleaved plasminogen activator inhibitor type 1 and a reactive-centre loop peptide. Structure*, **6**, 627–636.

Yan, Y., Li, Y., Munshi, S., Sardana, V., Cole, J. L., Sardana, M., Steinkuehler, C., Tomei, L., De Francesco, R., Kuo, L. C. & Chen, Z. (1998). *Complex of NS3 protease and NS4A peptide of BK strain hepatitis C virus: a 2.2 Å resolution structure in a hexagonal crystal form. Protein Sci.* **7**, 837–847.

Yang, J., Kloek, A. P., Goldberg, D. E. & Mathews, F. S. (1995). *The structure of Ascaris hemoglobin domain I at 2.2 Å resolution: molecular features of oxygen avidity. Proc. Natl Acad. Sci. USA*, **92**, 4224–4228.

Yang, X. & Moffat, K. (1995). *Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific Aspergillus ribotoxin, restrictocin. Structure*, **4**, 837–852.

Yao, N., Hesson, T., Cable, M., Hong, Z., Kwong, A. D., Le, H. V. & Weber, P. C. (1997). *Structure of the hepatitis C virus RNA helicase domain. Nature Struct. Biol.* **4**, 463–467.

Yee, V. C., Pedersen, L. C., LeTrong, I., Bishop, P. D., Stenkamp, R. E. & Teller, D. C. (1994). *Three-dimensional structure of a transglutaminase: human blood coagulation factor XIII. Proc. Natl Acad. Sci. USA*, **91**, 7296–7300.

Yeh, J. I., Claiborne, A. & Hol, W. G. J. (1996). *Structure of the native cystein-sulfenic acid redox center of enterococcal NADH peroxidase refined at 2.8 Å resolution. Biochemistry*, **35**, 9951–9957.

Yoshimoto, T., Kabashima, T., Uchikawa, K., Inoue, T., Tanaka, N., Nakamura, K. T., Tsuru, M. & Ito, K. (1999). *Crystal structure of prolyl aminopeptidase from Serratia marcescens. J. Biochem. (Tokyo)*, **126**, 559–565.

Zaitseva, I., Zaitsev, V., Card, G., Moshkov, K., Bax, B., Ralph, A. & Lindley, P. (1996). *The X-ray structure of human serum ceruloplasmin at 3.1 Å: nature of the copper centres. J. Biol. Inorg. Chem.* **1**, 15–23.

Zdanov, A., Schalk-Hihi, C., Menon, S., Moore, K. W. & Wlodawer, A. (1997). *Crystal structure of Epstein–Barr virus protein BCRF1, a homolog of cellular interleukin-10. J. Mol. Biol.* **268**, 460–467.

Zhang, A., Geisler, S. C., Smith, A. D., Resnick, D. A., Li, M. L., Wang, C. Y., Looney, D. J., Wong-Staal, F., Arnold, E. & Arnold, G. F. (1999). *A disulfide-bound HIV-1 V3 loop sequence on the surface of human rhinovirus 14 induces neutralizing responses against HIV-1. J. Biol. Chem.* **380**, 365–374.

Zhang, H., Gao, Y. G., van der Marel, G. A., van Boom, J. H. & Wang, A. H. (1993). *Simultaneous incorporations of two anticancer drugs into DNA. The structures of formaldehyde-cross-linked adducts of daunorubicin-d(CG(araC)GCG) and doxorubicin-d(CA(araC)GTG) complexes at high resolution. J. Biol. Chem.* **268**, 10095–10101.

Zhang, R., Evans, G., Rotella, F. J., Westbrook, E. M., Beno, D., Huberman, E., Joachimiak, A. & Collart, F. R. (1999). *Characteristics and crystal structure of bacterial inosine-5′-monophosphate dehydrogenase. Biochemistry*, **38**, 4691–4700.

Zhang, R. G., Scott, D. L., Westbrook, M. L., Nance, S., Spangler, B. D., Shipley, G. G. & Westbrook, E. M. (1995). *The three-dimensional crystal structure of cholera toxin. J. Mol. Biol.* **251**, 563–573.

Zhang, X., Morera, S., Bates, P. A., Whitehead, P. C., Coffer, A. I., Hainbucher, K., Nash, R. A., Sternberg, M. J., Lindahl, T. & Freemont, P. S. (1998). *Structure of an XRCC1 BRCT domain: a new protein–protein interaction module. EMBO J.* **17**, 6404–6411.

Zhu, X., Kim, J. L., Newcomb, J. R., Rose, P. E., Stover, D. R., Toledo, L. M., Zhao, H. & Morgenstern, K. A. (1999). *Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. Struct. Fold. Des.* **7**, 651–661.

Zuccola, H. J., Rozzelle, J. E., Lemon, S. M., Erickson, B. W. & Hogle, J. M. (1998). *Structural basis of the oligomerization of hepatitis delta antigen. Structure*, **6**, 821–830.

International Tables for Crystallography (2012). Vol. F, Chapter 1.4, pp. 39–44.

Chapter 1.4. Perspectives for the future

E. Arnold, M. G. Rossmann, D. M. Himmel, J. C. H. Spence and S. Sun

1.4.1. Gazing into the crystal ball (E. Arnold)

We live in an era when there are many wonderful opportunities for reaching new vistas of human experience. Some of the dreams that we hold may be achieved through scientific progress. Among the things we have learned in scientific research is to expect the unexpected – the crystal ball holds many surprises for us in the future. Alchemists of ancient times laboured to turn ordinary materials into precious metals. These days, scientists can create substances far more precious than gold by discovering new medicines and materials for high-technology industries. Crystallography has played an important role in helping to advance science and the human endeavour in the twentieth century. I expect to see great contributions from crystallography also in the twenty-first century.

Science of the twentieth century has yielded a great deal of insight into the workings of the natural world. Systematic advances are permitting dissection of the molecular anatomy of living systems. This has propelled us into a world where these insights can be brought to bear on problems of design. The impact in such fields as health and medicine, materials science, and microelectronics will be continually greater.

1.4.1.1. What can we expect to see in the future of science and technology in general?

Just as few were successful in predicting the ubiquitous impact of the Internet, it is difficult to predict which specific technologies will accomplish the transition into the culture of the future. It is possible to envision instantaneous telecommunication and videoconferencing with colleagues and friends throughout the world – anytime, anywhere – using small, portable devices. Access to computer-based information *via* media such as the Internet will become continually more facile and powerful. This will permit access to the storehouse of human knowledge in unprecedented ways, catalysing more rapid development of new ideas.

Experimental tests of new ideas will continue to play a crucial role in the guidance of scientific knowledge and reasoning. However, more powerful computing resources may change paradigms in which ever more powerful simulation techniques can bootstrap from primitive ideas to full-blown theories. I still expect that experiment will be necessary for the foreseeable future, since nearly every well designed experiment yields unexpected results, often at a number of levels.

In the realm of biology, greater understanding of the structure and mechanism of living processes will permit unprecedented advances in health and medicine. Even those scientists most sceptical of molecular-design possibilities would be likely to admit that revolutionary advances have been achieved. In the area of drug design, for example, structure-based approaches have yielded some of the most important new molecules currently being introduced worldwide for the treatment of human diseases ranging from AIDS and influenza to cancer and heart disease. This is a relatively young and very rapidly changing area, and it is reasonable to expect that we have witnessed only the tip of the iceberg. Dream drugs to control growth and form, ageing, intelligence, and other physiologically linked aspects of health and well-being may be developed in our lifetime. As greater understanding of the structural basis of immunogenicity emerges, we should also expect to benefit from structure-based approaches to vaccine design.

Other areas where molecular design will play revolutionary roles include the broad field of material sciences. Traditionally, ‘materials science’ referred to the development of materials with desired physical properties – strength, flexibility, and resistance to damage by physical and chemical agents. Now, materials science includes key foci in development of new biomaterials and in the burgeoning field of nanotechnology. The acrobatics of new smart materials could include computation at speeds that may be much faster than can be accomplished with silicon-based materials.

1.4.1.2. How will crystallography change in the future?

Potential future advances in the fields of crystallography, structural chemistry and biology are tantalizing. Owing to the limitations of current physical theories and experimental possibilities, large numbers of molecules have generally been required for detailed investigations of molecular phenomena. Given the complexity of large biological molecules, only techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) have been suitable for describing the detailed atomic structures of these systems. New technologies, unforseen and currently under development, will make the successful imaging of single specimens and single molecules at high resolution far more commonplace than is now achievable. Even now, we are beginning to see X-ray and electron microscopy (EM) methods harnessed to visualize single particles at high resolution. Methods such as coherent X-ray diffraction microscopy (CXDM), discussed in Chapter 9.3 of this volume, and the X-ray (free-electron) laser (XFEL) are challenging the limits of our ability to image single particles directly at high resolution (see Section 1.4.4, below). Merging of information from multiple specimens, as is currently done in electron microscopy, may be very powerful in X-ray microscopy as well.

X-ray sources will continue to evolve. High-intensity synchrotron sources have allowed the development of dramatically faster and higher-quality diffraction data measurements. Complete multiwavelength X-ray diffraction data sets have been measured from frozen protein crystals containing selenomethionine (SeMet) in less than one hour, leading to nearly automatic structure solution by the multiwavelength anomalous diffraction (MAD) technique. At present, such synchrotron facilities are enormous national or multinational facilities that sap the electric power of an entire region. Perhaps portable ‘table-top’ X-ray sources will be developed that can be used to create synchrotron-like intensity in the laboratory. If such sources could have a tunable energy or wavelength, then experiments such as MAD would be routinely accessible within the laboratory. Better time resolution of molecular motion and of chemical reactions will be achieved with higher-intensity sources.

Sample preparation for macromolecular structural studies has undergone a complete revolution thanks to the advent of

Copyright © 2012 International Union of Crystallography

recombinant DNA methods. Early macromolecular studies were limited to materials present in large abundance. By the late 1970s and early 1980s, molecular biology made it possible to obtain desired gene products in large amounts, and new methods of chemical synthesis permitted production of large quantities of defined oligonucleotides. Initial drafts of the entire human genome have been mapped and sequenced, allowing even broader access to genes for study. The combination of structural genomics and already ongoing studies will lead to knowing the structure of the entire human proteome in a finite amount of time. Many materials are still challenging to produce in quantities sufficient for structural studies. Engineering methods (site-directed mutagenesis, combinatorial mutagenesis and directed evolution techniques) have permitted additional sampling of molecular diversity, and we can expect that even more powerful methods will be developed. Engineering of solubility and crystallizability will help make more problems tractable for study. Perhaps, as suggested before, techniques for visualization of single molecules may become adequate for *in situ* visualization of molecular interactions in living cells and organisms. However, traditional considerations of amount, purity, specific activity *etc.* will remain important, as will hard work and good luck.

The phase problem has continued to be a stumbling block for structure determination. Experimental methods, including isomorphous replacement and anomalous-scattering measurements, are currently a mainstay, and will be for the foreseeable future. New isomorphous heavy-atom reagents and preparation methods will emerge; witness the valuable engineering of derivatives *via* mutagenesis to add such heavy-atom-binding sites as cysteine residues in proteins. Anomalous-scattering measurements from macromolecular crystals containing heavy metals or SeMet replacements of methionine residues in proteins have led to tremendous acceleration of the phase-problem solution for many structures – especially in the last two decades with the availability of 'tunable' synchrotron radiation. It should become possible to take better advantage of anomalous-scattering effects from lighter atoms already present in biological molecules: sulfur, phosphorus, oxygen, nitrogen and carbon. The increasingly higher intensity of synchrotron-radiation sources may permit structure solution from microcrystals of macromolecules. Incorporation of non-standard amino acids into proteins will become more common, leading to a vast array of new substitution possibilities. Molecular-replacement phasing from similar structures or from noncrystallographically redundant data is now commonplace and is continuing to become easier. Systematic molecular replacement using all known structures from databases may prove surprisingly powerful, if we can learn how to position small molecular fragments reliably. Systematic molecular-replacement approaches should help identify what folds may be present in a crystalline protein of unknown structure. Direct computational assaults on the phase problem are also becoming more aggressive and successful, although direct-methods approaches still work best for small macromolecular structures with very high resolution data.

Crystallography and structural biology have been helping to drive advances in three-dimensional visualization technology. Versatile molecular-graphics packages have been among the most important software applications for the best three-dimensional graphics workstations. Now that personal computers are being mass-produced with similar graphics capabilities, we are beginning to see a molecular-graphics workstation at every computer, whether desktop or portable (terms that soon may become antiquated since everything will become more compact). Modes of input will include direct access to thought processes, and computer output devices will extend beyond light and sound. Universal Internet access will provide immediate access to the rapidly increasing store of molecular information. As a result, we will achieve a more thorough understanding of patterns present in macromolecular structures: common folds of proteins and nucleic acids, three-dimensional motifs, and evolutionary relationships among molecules. Simulations of complex molecular motions and interactions will be easier to display, making movies of molecules in motion commonplace. Facile 'virtual reality' representation of molecules will be a powerful research and teaching aid. Chemical reaction mechanisms will become better understood over time through interplay among theory, experiment and simulation. The ability to simulate all coupled chemical reactions in living cells and organisms will be achieved over time.

Advances in computational productivity depend on the intricate co-evolution of hardware and software. For silicon-based transistor chips, raw computational speed doubles approximately every 18 months (Moore's law). Tools and software for writing software will continue to advance rapidly. With greater modularity of software tools, it will become easier to coordinate existing programs and program suites. Enhanced automation, parallelization and development of new algorithms will also increase speed and throughput. More powerful software heuristics involving artificial intelligence, expert systems, neural nets and the like may permit unexpected advances in our understanding of the natural world.

In summary, we eagerly await what the future of science and of studies of molecular structure will bring. There is every reason to expect the unexpected. If the past is a guide, many new flowers will bloom to colour our world in bright new ways.

1.4.2. Brief comments on *Gazing into the crystal ball*
(M. G. Rossmann)

Eddy Arnold and I had planned to write a joint commentary about our vision of the future of macromolecular crystallography. However, when Eddy produced the first draft of '*Perspectives for the future*', I was fascinated by his wide vision. I felt it more appropriate and far more interesting to make my own brief comments, stimulated by Eddy's observations.

When I was a graduate student in Scotland in the 1950s, physics departments were called departments of 'Natural Philosophy'. Clearly, the original hope had been that some aspects of science were all encompassing and gave insight to every aspect of observations of natural phenomena. However, in the twentieth century, with rapidly increasing technological advances, it appeared to be more and more difficult for any one person to study all of science. Disciplines were progressively subdivided. Learning became increasingly specialized. *International Tables* were created, and updated, for the use of a highly specialized and small community of crystallographic experts.

As I read Eddy's draft article, I became fascinated by the wide impact he envisioned for crystallography in the next few decades. Indeed, the lay person, reading his article, would barely be aware that this was an article anticipating the future impact of crystallography. The average reader would think that the topic was the total impact of science on our civilization. Thus, to my delight, I saw that crystallography might now be a catalyst for the reunification of fragmented science into a coherent whole. I therefore hope that these new *Tables* commissioned by the International Union of Crystallography may turn out to be a significant help to further the trend implied in Eddy's article.

1.4.3. Additional comments on *Gazing into the crystal ball*
(D. M. Himmel)

I wish to expand briefly on Eddy's and Michael's comments to entertain the question of how macromolecular crystallography and structural biology might change the future. New detector technologies, such as shutterless detector systems that can collect data in continuous rotation mode (see, for example, Miyoshi *et al.*, 2008), are now being developed and deployed at synchrotrons. These detectors will improve the quality of data collection and integration. It is possible that the emergence of ever more rapid and more sensitive X-ray detectors, along with new X-ray technologies, will lead to multiple-wavelength rapidly pulsed X-ray beams that are synchronized by wavelength with the detector. At each crystal rotation position, the beam will rapidly run through each of the several tuned wavelengths and record a diffraction pattern for each wavelength. This sort of setup can be used to collect MAD data from a single data pass through the Ewald sphere. In cases where crystals are particularly sensitive to X-ray damage, this approach, along with the ability to detect anomalous signals from ever-smaller elements in the Periodic Table, may make MAD more competitive with molecular replacement as the method of choice for structure determination. Molecular replacement itself will become more powerful as it is combined with *ab initio* phasing and computational chemistry methods. In the future, due to other emerging technologies such as described in Section 1.4.4 below, whether or not one can crystallize a macromolecule may no longer limit one's ability to determine a structure by X-ray diffraction. Here, I will focus on two additional points: (1) how X-ray crystallography and allied fields (such as NMR and EM) have already started to transform the physical sciences, and (2) how they might transform the overall human experience as well.

In the mid-twentieth century, the physical sciences to a considerable extent were divided into separate fiefdoms that often competed with each other rather than fostered a collaborative spirit. Physics, chemistry, virology and biochemistry, to name a few, each attempted to stand on its own, each using its own independent preferred jargon and preferred explanations for scientific questions of the day. By contrast, what we see happening today is a convergence of all these once-separate fields to understand first principles right down to the molecular level and beyond. More than that, macromolecular crystallography and allied fields are playing a substantial role in catalysing this convergence. The determination and subsequent analysis of a macromolecular structure (such as a protein in complex with nucleic acid, co-factors and/or small-molecule ligands) of necessity culminates from the application of physics, mathematics and chemistry to begin with, followed by various biological sciences to understand the context of a molecular structure. But for physicists, chemists and biologists to speak to each other productively, they increasingly must share the same jargon and learn each other's disciplines. Sir Isaac Newton once wrote, 'If I have seen further it is by standing on the shoulders of giants.' The giants upon which we stand today in crystallography were the leaders of those separate fields that today find their synthesis in the determination and understanding of molecular structure. It is reasonable to assume that this unifying spirit will continue and will foster greater breakthroughs in structure determination and understanding the properties of molecules in the real world.

In addition to bringing various disciplines of science together, macromolecular crystallography and allied fields are likely to revolutionize the way people live in the future. Richard Feynman was reported to have said, 'What I cannot create, I do not understand.' Chapters 3.1, 3.2, and 4.3 of this volume describe the harnessing of gene expression and protein engineering to further the aims of the X-ray crystallographic experiment. Macromolecular engineering, however, does not stop there. Recent years have seen a rising interest in the field of nanotechnology, which, according to some definitions, aims to design machines and technologies that operate on a scale of about 100 nm or smaller (Farokhzad & Langer, 2009). These technologies generally encompass applications of our understanding of chemistry and physics. As this field matures, it will encounter many of the problems on the molecular level that have already been solved by the machines and devices that exist in biological systems, such as enzymes, molecular motors and structural proteins. Proofs of concept have recently been described in which components of biological molecules were redesigned for applications outside their usual environment (Goel & Vogel, 2008; van den Heuvel & Dekker, 2007; Lewis *et al.*, 2011), or in which principles learned from structural biology have been applied to the construction of completely artificial molecular devices (Ceroni *et al.*, 2010). The discoveries of structural biology may well light the way to the nanotechnologies of the future. These technologies, by operating at the molecular level like never before, will enable the manufacture of superior fabrics and materials, improve medical diagnostics, and revolutionize electronics and photonics in a whole host of devices, from computers and robotics to communication and the efficient harnessing of unconventional energy sources like light. Chemical catalysis will be performed in industry with such tight control that there will be far fewer unwanted side reactions than is commonplace today, so that, for example, medicines and other materials can be manufactured cheaply to an exceptionally high level of purity. Drug-delivery vehicles will be developed that precisely target a cell type, tissue type (Farokhzad & Langer, 2009) or even a pathogen. Some of these molecular vehicles will be modelled on methods used by viruses such as influenza, HIV or even bacteriophages, using a harpoon or plasma membrane fusion strategy, and will release their cargo in response to a chemical or other trigger. Others will employ active transport, in which the therapeutic agent will be guided to its target by remote control and carry diagnostics modules for the ride. The development of these nanotechnologies will both benefit from nanotechnological spin-offs of structural studies as well as require X-ray crystallography and other structural techniques to aid in the analysis of the nanomachines.

To peer at an image of individual molecules at atomic or near-atomic resolution, X-ray crystallography has served to provide the missing lens needed to focus that image. New and maturing fields described in the following sections will supply other ways to provide that missing lens, or, indeed, to obtain the initial phase information needed to determine a higher-resolution X-ray crystal structure.

1.4.4. *Gazing into the crystal ball* – the X-ray free-electron laser
(J. C. H. Spence)

The recent invention of the pulsed hard X-ray (free-electron) laser (XFEL) is certain to impact structural biology, particularly in the areas of protein nanocrystal analysis (Chapman *et al.*, 2011), single-particle imaging (Siebert *et al.*, 2011), time-resolved crystallography and solution scattering (see the forthcoming reviews in *Reports on Progress in Physics* by Spence and Chapman). Current hard-X-ray machines provide about 10^{12}

photons in each 40 femtosecond pulse, and are capable of reading out perhaps 120 of the resulting diffraction-pattern 'snapshots' every second. Such a beam, focused to micron dimensions, vaporizes the sample, but it has been discovered that a useful pattern is obtained before radiation damage commences (due to the photoelectron cascade). The method has given 2 Å resolution data from micron-sized protein nanocrystals, and, if sufficiently brief pulses are used, allows about 100 times greater dose to be delivered than the Henderson 'safe dose' (see Chapter 10.3). The snapshot data consist of partial reflections. As a consequence of the fully coherent nature of the radiation, for the smallest submicron nanocrystals the data show interference fringes between the Bragg reflections that facilitate iterative phasing (Spence *et al.*, 2011). Sample delivery has been based on a continuously flowing liquid jet of micron or submicron dimensions, freely flowing in vacuum, with gas focusing at a nozzle to prevent clogging. Merging of millions of nanocrystal snapshots to obtain full reflections has created new challenges for data analysis, as has the development of MAD phasing for the time-resolved absorption involved.

Diffraction patterns may also be obtained from single particles such as a virus or whole cell, commonly injected from a nebulizer in a gas-focused stream into vacuum. Each snapshot gives one projection of the particle in a random orientation, so that three-dimensional reconstruction requires a solution to the difficult problems of orientation determination and phasing of single-particle diffraction patterns. The available X-ray fluence per shot, together with the minimum amount of scattering needed for orientation determination, has so far limited resolution to about 30 nm. However, more powerful X-ray lasers, smaller focused spots and improvements in 'hit rate' are bound to improve resolution to the predicted 1 nm resolution limit for particles too thick for study by cryogenic electron microscopy (cryo-EM). Scattering in the water window, around 500 eV, gives greatly increased protein/water contrast, but wavelength-limited resolution. Conformational variability imposes similar limitations to those encountered in cryo-EM, and the merging of multiple projections from similar cells remains an important challenge for the future, perhaps based on topological constraints. The XFEL also offers unprecedented opportunities for time-resolved imaging, spanning the range from the femtosecond timescale important for electron-transfer reactions in biochemistry, to the slower microsecond processes of protein activity. In favourable cases, a fast optical trigger exists for pump–probe studies, while in others chemical reactions (such as the enzyme cycle) might be followed in mixed and flowing solutions. For this purpose, the correlated fluctuations in 'snapshot' small-angle X-ray (SAX) patterns may prove useful, since they offer a hit rate of 100%. Since these patterns are two-dimensional for particles frozen in space or time, they contain more information than conventional one-dimensional SAX patterns. The ability to reconstruct an image of one particle using the scattering from many randomly oriented particles frozen in space (without modelling) has recently been demonstrated (Saldin *et al.*, 2011). In summary, the XFEL has opened up many new exciting possibilities for structural and dynamic biology, based on entirely new experimental arrangements (now far from optimized) and offering great scope for developments in this highly interdisciplinary field, which spans laser, detector and particle injector physics, diffraction physics, and structural biology. We anticipate rapid progress in methods for the growth of suitable protein microcrystals, especially for membrane proteins. The XFEL is then certain to provide a wealth of new information on molecular mechanisms in biology, as techniques are refined and more powerful X-ray lasers are constructed.

1.4.5. Electron microscopy's impact on structural biology (S. Sun)

Just like crystallography half a century ago, cryo-electron microscopy (cryo-EM) has been developing rapidly and has become one of the dominant techniques in structural biology. Ever since an electron microscope was first constructed, there has been a desire to examine biological samples. There were two major obstacles to this task, one being how to obtain a sample thin enough for the electron beam to go through, the other being the preservation of the sample in the high vacuum of the microscope. Early EM studies of biological specimens combined fixation, dehydration, embedding and sectioning with application of heavy metals to provide contrast in tissue samples. These methods cannot preserve the biological samples in their native state, but merely produce an outline rather than provide intricate three-dimensional structural details. In the 1980s, Dr Jacques Dubochet and his co-workers at the EMBL developed a method of producing vitrified biological samples that made it possible to study biological specimens in their native state (Dubochet *et al.*, 1981, 1988). Together with the invention of high-stability cold stages and transfer mechanisms for electron microscopes, there has been an explosion of studies using cryo-EM for biological samples (Frank, 1989; Ruiz *et al.*, 1994; Frank, 2009; Wade & Hewat, 1994; Ben-Harush *et al.*, 2010).

Single-particle cryo-EM and cryo-electron tomography (cryo-ET) are the most commonly used techniques for the study of biological samples. Owing to the sensitivity to radiation that causes structural damage, only small electron doses can be used to examine biological samples. As a result, the signal-to-noise ratio is low for cryo-EM and especially for cryo-ET. In single-particle cryo-EM, this can be compensated for by collecting two-dimensional (2D) projections of a large number of particles and averaging projections representing the same orientation. Viruses have been among the most popular subjects for single-particle cryo-EM reconstruction because of their high symmetry. With several hundred particles, a 20 Å or so resolution reconstruction can be obtained with relative ease. For particles with low or no symmetry, more data need to be collected to achieve the same resolution. Two of the better-studied asymmetric particles by single particle cryo-EM are the ribosome and GroEL (Frank, 2001; Roseman *et al.*, 2001).

Although the theoretically achievable resolution limit using an electron microscope should be related to the wavelength of the electron beam, in reality there are many factors that affect the resolution, some of which are microscope-related or sample-related. These factors include, but are not limited to, the quality of the microscope, the stability of the sample towards radiation damage, and sample or beam movements caused by electric or magnetic field variation, temperature changes, or mechanical vibrations. It is also essential that the individual samples are sufficiently homogeneous to allow meaningful averaging within the desired resolution. The quality of a specific cryo-EM experimental data set can be judged by the 'Thon rings', provided by the averaged radial Fourier transformation of the particle images. These rings show the averaged resolution limit of the image signal and thus show how far it might be possible to extend the resolution, provided that all relevant parameters are accurately determined. Once the best possible data have been collected on a particular sample, numerous computational steps follow to achieve a good result. These include boxing (selecting particles within a 'box' and normalizing the density to a standard variation), contrast-transfer-function correction, 2D alignment and 3D reconstruction. Many independent groups have developed various software packages for 3D EM reconstruction such as *SPIDER* (see Chapter 19.8), *IMAGIC 4D* (Chapter 19.9), *EMAN* (see Chapter 19.10), *AUTO3DEM* and *FREALIGN* (Ludtke *et al.*, 1999; Shaikh *et al.*, 2008; Yan *et al.*, 2007; Grigorieff, 2007). Numerous aspects of the reconstruction procedure are still being developed to achieve higher resolution, such as individual-particle contrast-transfer-function correction, magnification correction, local refinement in 2D alignment, and averaging among quasi-equivalent subunits within the icosahedral asymmetric unit after the 3D reconstruction. The highest resolution structure obtained for a virus is the 3.1 Å resolution structure of the cytoplasmic polyhedrosis virus.

The definition of resolution of a single-particle cryo-EM reconstruction is different from that used for assessing a crystal structure. The resolution of an EM reconstruction is usually estimated using the 'Fourier shell correlation' method (van Heel & Schatz, 2005). This requires dividing the randomly chosen images into two separate groups and computing reconstructions assuming the previously determined orientations. The resultant maps are then Fourier transformed to permit a comparison of the Fourier amplitudes and phases. The resolution at which the correlation between the amplitudes falls below 0.5 or the average phase difference is greater than 45° is usually a conservative estimate of resolution used by many groups, although other definitions are also in use (Henderson, 2004). Just as is the case for crystallographic maps, the quality of maps at the same resolution can vary depending on the degree of accuracy of the phase determination. Therefore, a visual inspection of a 3D EM map should serve as a useful check.

In contrast to single-particle cryo-EM, cryo-ET can be used to study biological samples that lack homogeneity (Milne & Subramaniam, 2009; Li & Jensen, 2009; Hoenger & McIntosh, 2009; Subramaniam *et al.*, 2007). In this technique, images of the same sample oriented by a succession of tilt angles are taken about a common axis and are subsequently combined into a three-dimensional map. Because the same sample is repeatedly exposed to the electron beam, the allowable electron dose has to be spread over all the tilted images, making the exposure for each image much less than for single-particle cryo-EM. This significantly reduces the limit of achievable resolution. In addition, the physical limitation on the tilt angles results in a 'missing wedge' problem. Dual-axis tomography has been developed to alleviate this problem to some extent. Averaging of the homogeneous portion of the heterogeneous sample has been used when appropriate to increase the resolution of a cryo-ET experiment (Bartesaghi & Subramaniam, 2009).

Cryo-EM methods have been an important addition to the family of structural biology tools. With single-particle cryo-EM, near-atomic-resolution structures can be obtained without the requirement of crystal formation, which is still the rate-limiting step for crystallographic studies. Cryo-ET is particularly useful in studying large molecular complexes and cells that do not possess any symmetry. Further improvements, such as the use of phase plates, further development of automated data-collection software and perfection of reconstruction algorithms, make it likely that cryo-EM will overtake crystallography or NMR as the primary tool for structural biology at low to medium resolution. However, crystallography, NMR and other tools will still be essential to obtain detailed, atomic resolution information.

References

Bartesaghi, A. & Subramaniam, S. (2009). *Membrane protein structure determination using cryo-electron tomography and 3D image averaging. Curr. Opin. Struct. Biol.* **19**, 402–407.

Ben-Harush, K., Maimon, T., Patla, I., Villa, E. & Medalia, O. (2010). *Visualizing cellular processes at the molecular level by cryo-electron tomography. J. Cell Sci.* **123**, 7–12.

Ceroni, P., Credi, A., Venturi, M. & Balzani, V. (2010). *Light-powered molecular devices and machines. Photochem. Photobiol. Sci.* **9**, 1561–1573.

Chapman, H. N., Fromme, P., Barty, A., White, T. A., Kirian, R. A., Aquila, A., Hunter, M. S., Schulz, J., DePonte, D. P., Weierstall, U., Doak, R. B., Maia, F. R., Martin, A. V., Schlichting, I., Lomb, L., Coppola, N., Shoeman, R. L., Epp, S. W., Hartmann, R., Rolles, D., Rudenko, A., Foucar, L., Kimmel, N., Weidenspointner, G., Holl, P., Liang, M., Barthelmess, M., Caleman, C., Boutet, S., Bogan, M. J., Krzywinski, J., Bostedt, C., Bajt, S., Gumprecht, L., Rudek, B., Erk, B., Schmidt, C., Hömke, A., Reich, C., Pietschner, D., Strüder, L., Hauser, G., Gorke, H., Ullrich, J., Herrmann, S., Schaller, G., Schopper, F., Soltau, H., Kühnel, K. U., Messerschmidt, M., Bozek, J. D., Hau-Riege, S. P., Frank, M., Hampton, C. Y., Sierra, R. G., Starodub, D., Williams, G. J., Hajdu, J., Timneanu, N., Seibert, M. M., Andreasson, J., Rocker, A., Jönsson, O., Svenda, M., Stern, S., Nass, K., Andritschke, R., Schröter, C. D., Krasniqi, F., Bott, M., Schmidt, K. E., Wang, X., Grotjohann, I., Holton, J. M., Barends, T. R., Neutze, R., Marchesini, S., Fromme, R., Schorb, S., Rupp, D., Adolph, M., Gorkhover, T., Andersson, I., Hirsemann, H., Potdevin, G., Graafsma, H., Nilsson, B. & Spence, J. C. (2011). *Femtosecond X-ray protein nanocrystallography. Nature (London)*, **470**, 73–77.

Dubochet, J., Adrian, M., Chang, J. J., Homo, J. C., Lepault, J., McDowall, A. W. & Schultz, P. (1988). *Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys.* **21**, 129–228.

Dubochet, J., Booy, F. P., Freeman, R., Jones, A. V. & Walter, C. A. (1981). *Low temperature electron microscopy. Annu. Rev. Biophys. Bioeng.* **10**, 133–149.

Farokhzad, O. C. & Langer, R. (2009). *Impact of nanotechnology on drug delivery. ACS Nano*, **3**, 16–20.

Frank, J. (1989). *Image analysis of single macromolecules. Electron Microsc Rev.* **2**, 53–74.

Frank, J. (2001). *Cryo-electron microscopy as an investigative tool: the ribosome as an example. Bioessays*, **23**, 725–732.

Frank, J. (2009). *Single-particle reconstruction of biological macromolecules in electron microscopy – 30 years. Q. Rev. Biophys.* **42**, 139–158.

Goel, A. & Vogel, V. (2008). *Harnessing biological motors to engineer systems for nanoscale transport and assembly. Nat. Nanotechnol.* **3**, 465–475.

Grigorieff, N. (2007). *FREALIGN: high-resolution refinement of single particle structures. J. Struct. Biol.* **157**, 117–125.

Heel, M. van & Schatz, M. (2005). *Fourier shell correlation threshold criteria. J. Struct. Biol.* **151**, 250–262.

Henderson, R. (2004). *Realizing the potential of electron cryo-microscopy. Q. Rev. Biophys.* **37**, 3–13.

Heuvel, M. G. van den & Dekker, C. (2007). *Motor proteins at work for nanotechnology. Science*, **317**, 333–336.

Hoenger, A. & McIntosh, J. R. (2009). *Probing the macromolecular organization of cells by electron tomography. Curr. Opin. Cell Biol.* **21**, 89–96.

Lewis, J. C., Coelho, P. S. & Arnold, F. H. (2011). *Enzymatic functionalization of carbon–hydrogen bonds. Chem. Soc. Rev.* **40**, 2003–2021.

Li, Z. & Jensen, G. J. (2009). *Electron cryotomography: a new view into microbial ultrastructure. Curr. Opin. Microbiol.* **12**, 333–340.

Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). *EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol.* **128**, 82–97.

Milne, J. L. & Subramaniam, S. (2009). *Cryo-electron tomography of bacteria: progress, challenges and future prospects. Nat. Rev. Microbiol.* **7**, 666–675.

Miyoshi, T., Igarashi, N., Matsugaki, N., Yamada, Y., Hirano, K., Hyodo, K., Tanioka, K., Egami, N., Namba, M., Kubota, M., Kawai, T. & Wakatsuki, S. (2008). *Development of an X-ray HARP-FEA detector system for high-throughput protein crystallography. J. Synchrotron Rad.* **15**, 281–284.

Roseman, A. M., Ranson, N. A., Gowen, B., Fuller, S. D. & Saibil, H. R. (2001). *Structures of unliganded and ATP-bound states of the Escherichia coli chaperonin GroEL by cryoelectron microscopy*. *J. Struct. Biol.* **135**, 115–125.

Ruiz, T., Erk, I. & Lepault, J. (1994). *Electron cryo-microscopy of vitrified biological specimens: towards high spatial and temporal resolution*. *Biol. Cell*, **80**, 203–210.

Saldin, D. K., Poon, H. C., Bogan, M. J., Marchesini, S., Shapiro, D. A., Kirian, R. A., Weierstall, U. & Spence, J. C. H. (2011). *New light on disordered ensembles: Ab-initio structure determination of one particle from scattering fluctuations of many copies*. *Phys. Rev. Lett.* **106**, 115501.

Shaikh, T. R., Gao, H., Baxter, W. T., Asturias, F. J., Boisset, N., Leith, A. & Frank, J. (2008). *SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs*. *Nat. Protoc.* **3**, 1941–1974.

Siebert, M. M. *et al.* (2011). *Single mimivirus particles intercepted and imaged with an X-ray laser*. *Nature (London)*, **470**, 78–81.

Spence, J. C., Kirian, R. A., Wang, X., Weierstall, U., Schmidt, K. E., White, T., Barty, A., Chapman, H. N., Marchesini, S. & Holton, J. (2011). *Phasing of coherent femtosecond X-ray diffraction from size-varying nanocrystals*. *Opt. Express*, **19**, 2866–2873.

Subramaniam, S., Bartesaghi, A., Liu, J., Bennett, A. E. & Sougrat, R. (2007). *Electron tomography of viruses*. *Curr. Opin. Struct. Biol.* **17**, 596–602.

Wade, R. H. & Hewat, E. A. (1994). *Cryoelectron microscopy of macromolecular complexes*. *Biol. Cell*, **80**, 211–220.

Yan, X., Sinkovits, R. S. & Baker, T. S. (2007). *AUTO3DEM – an automated and high throughput program for image reconstruction of icosahedral particles*. *J. Struct. Biol.* **157**, 73–82.

International Tables for Crystallography (2012). Vol. F, Chapter 2.1, pp. 45–63.

PART 2. BASIC CRYSTALLOGRAPHY

Chapter 2.1. Introduction to basic crystallography

J. Drenth

2.1.1. Crystals

It is always amazing to see how large molecules, such as proteins, nucleic acids and their complexes, order themselves so neatly in a crystalline arrangement. It is surprising because these large molecules have irregular surfaces with protrusions and cavities, and hydrophilic and hydrophobic spots. Nevertheless, they pack themselves into an orderly arrangement in crystals of millimetre sizes.

Crystals of biological macromolecules are, like most other crystals, not ideal. The X-ray diffraction pattern fades away at diffraction angles corresponding to lattice-plane distances between 1 and 2 Å or even worse. This is not so surprising, since protein crystals are relatively soft. The interaction energy between protein molecules in crystals is of the order of 63×10^{-21} J per protein molecule, or approximately 15 kT (Haas & Drenth, 1995). This corresponds to about ten hydrogen bonds, four salt bridges, or a 400 Å^2 buried hydrophobic surface. Although this energy might not be very different from crystalline interactions between small molecules, the large size of the protein molecules or macromolecular assemblies makes the crystals much more sensitive to distorting forces. Irregularities in the crystal lattice can also stem from the incorporation of impurities – either foreign substances or slightly denatured molecules from the parent protein. Moreover, some molecules may be incorrectly oriented, because the difference in interaction energy between different orientations is rather small. Also, amino-acid side chains assume more than one conformation. These are static irregularities. In addition, dynamic disorder exists: parts of the macromolecule are flexible and affect the X-ray diffraction pattern just as the temperature does.

By neglecting distortions caused by lattice imperfections, crystals are found to have a repeating unit, the unit cell, with basis vectors **a**, **b** and **c**, and angles α, β and γ between them (Fig. 2.1.1.1). The enormous number of unit cells in a crystal are stacked in three dimensions, in an orderly way, with the origins of the unit cells forming a grid or lattice. In Fig. 2.1.1.2, part of a crystalline lattice containing 5 × 3 × 3 unit cells is drawn.

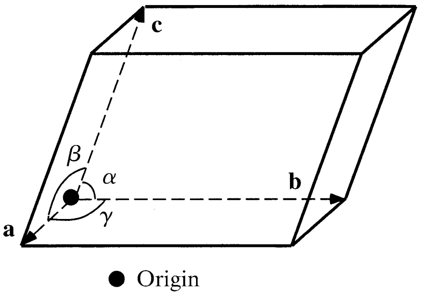


Figure 2.1.1.1
One unit cell with axes **a**, **b** and **c**. The angles between the axes are α, β and γ. Note that the axial system is right-handed. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

It is customary to call the direction along the unit-cell vector **a** the x direction in the lattice; similarly, y is along **b**, and z along **c**.

Crystallographers use a simple system to indicate the planes in a crystal lattice. For instance, the plane containing the unit-cell vectors **a** and **b** is called (001), and the plane containing the vectors **b** and **c** is called (100). The plane (010) contains the vectors **a** and **c**. It should be pointed out that these planes are not limited to one unit cell, but extend through the entire crystal. Moreover, each of these three planes is only one member of a set of parallel and equidistant planes: the set (001), the set (100) and the set (010). For each set, the lattice planes pass through all lattice points, where the lattice points are at the corners of the unit cells (see Fig. 2.1.1.2). Besides the sets of planes (001), (100) and (010), many more sets of parallel and equidistant planes can be drawn through the lattice points. In Fig. 2.1.1.3, this is done for a two-dimensional lattice. Lattice planes always divide the unit-cell vectors **a**, **b** and **c** into a number of equal parts. If the lattice planes divide the **a** vector of the unit cell into h equal parts, the first index for this set of planes is h. The second index, k, is related to the division of **b**, and the third index, l, to the division of **c**. If the set of lattice planes is parallel to a basis unit-cell vector, the corresponding index is 0. Indices for lattice planes are given in parentheses. They should not be confused with directions of vectors connecting lattice points; these are given in square brackets: [uvw], where u is the coordinate in the **a** direction expressed as the number of **a**'s, v in the **b** direction expressed as the number of **b**'s and w in the **c** direction expressed as the number of **c**'s. u, v and w are taken as the simplest set of whole numbers. For instance, [100] is along **a**; [200] has the same direction, but [100] is used instead. [111] points from the origin to the opposite corner of the unit cell.

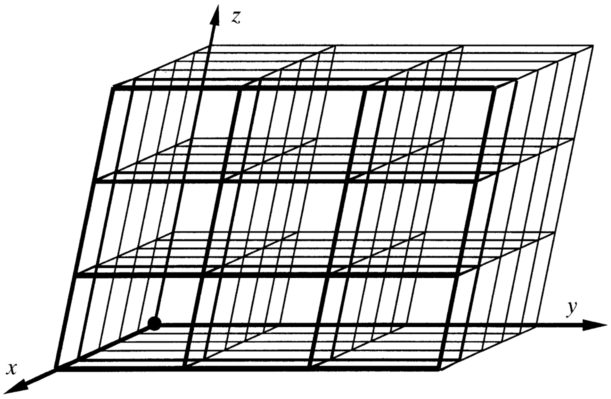


Figure 2.1.1.2
A set of 5 × 3 × 3 unit cells. The points where the lines intersect are called lattice points. The axes x and y form a (001) plane, which is one member of the set of parallel and equidistant (001) planes; y and z form a (100) plane, and z and x a (010) plane. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

Copyright © 2012 International Union of Crystallography

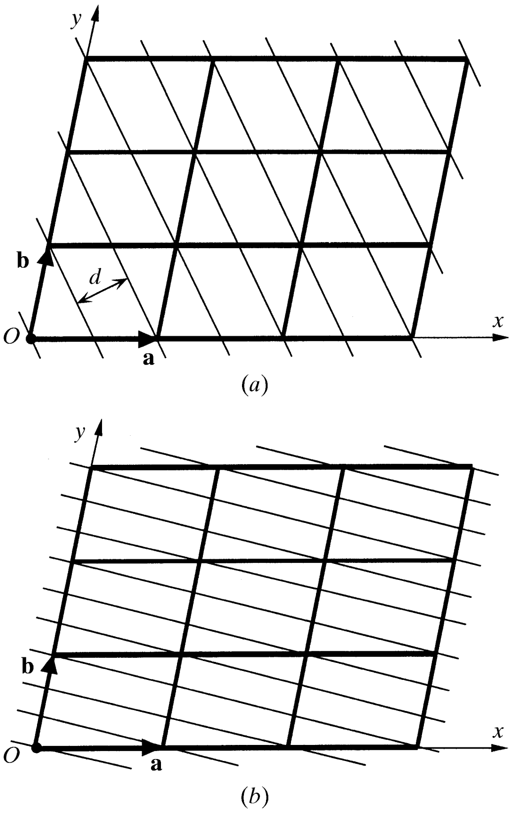


Figure 2.1.1.3

A two-dimensional lattice with 3 × 3 unit cells. In both (*a*) and (*b*), a set of equidistant parallel lattice planes is drawn. They pass through all lattice points. Lattice planes always divide the unit-cell axes into a whole number of equal parts – 1, 2, 3 *etc.* For instance, in (*a*), the vector **a** of the unit cell is cut into two parts, and the vector **b** into only one part. This set of planes is then given the indices $h = 2$ and $k = 1$. In three dimensions, there would be a third index, l. In (*b*), the set of lattice planes has the indices $h = 1$ and $k = 3$. In general, lattice planes have the indices (hkl), known as Miller indices. If a set of lattice planes is parallel to an axis, the corresponding index is 0. For instance, (001) is the set of planes parallel to the unit-cell vectors **a** and **b**. Note that the projection of $\mathbf{a}/h$ on the line normal to the lattice plane is equal to the lattice-plane distance d. This is also true for $\mathbf{b}/k$. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

The choice of the unit cell is not unique and, therefore, guidelines have been established for selecting the standard basis vectors and the origin. They are based on symmetry and metric considerations:

(1) The axial system should be right-handed.
(2) The basis vectors should coincide as much as possible with directions of highest symmetry.
(3) The cell taken should be the smallest one that satisfies condition (2).
(4) Of all lattice vectors, none is shorter than **a**.
(5) Of those not directed along **a**, none is shorter than **b**.
(6) Of those not lying in the *ab* plane, none is shorter than **c**.
(7) The three angles between the basis vectors **a**, **b** and **c** are either all acute (<90°) or all obtuse (≥90°).

It should be noted that the rules for choosing **a**, **b** and **c** are not always obeyed, because of other conventions (see Section 2.1.3). Condition (3) sometimes leads to a centred unit cell instead of a primitive cell. Primitive cells have only one lattice point per unit cell, whereas non-primitive cells contain two or more lattice points. They are designated *A*, *B* or *C* if opposite faces of the cell are centred: *A* for *bc* centring, *B* for *ac* centring and *C* for *ab* centring. If all faces are centred, the designation is *F*, and if the cell is body-centred, it is *I* (Fig. 2.1.1.4).

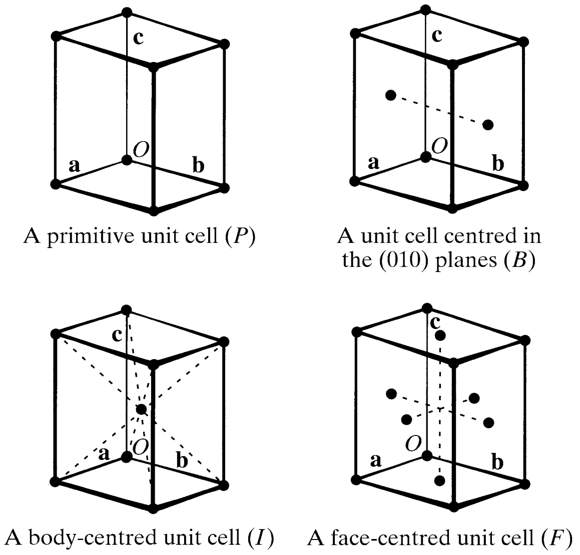

c
O
a
b
A primitive unit cell (P)
c
O
a
b
A unit cell centred in the (010) planes (B)
c
O
a
b
A body-centred unit cell (I)
c
O
a
b
A face-centred unit cell (F)

Figure 2.1.1.4

Non-centred and centred unit cells. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

2.1.2. Symmetry

A symmetry operation can be defined as an operation which, when applied, results in a structure indistinguishable from the original one. According to this definition, the periodic repetition along **a**, **b** and **c** represents translational symmetry.

In addition, rotational symmetry exists, but only rotational angles of 60, 90, 120, 180 and 360° are allowed (*i.e.* rotation over $360/n$ degrees, where n is an integer). These correspond to n-fold rotation axes, with $n = 6, 4, 3, 2$ and 1 (identity), respectively. Rotation axes with $n = 5$ or $n > 6$ are not found as crystallographic symmetry axes, because translations of unit cells containing these axes do not completely fill three-dimensional space. Another type of rotational symmetry axis is the screw axis. It combines a rotation with a translation. For a twofold screw axis, the translation is over 1/2 of the unit-cell length in the direction of the axis; for a threefold screw axis, it is 1/3 or 2/3 *etc.* In this way, the translational symmetry operators can be obeyed. The requirement that translations are 1/2, 1/3, 2/3 *etc.* of the unit-cell length does not exist for individual objects that are not related by crystallographic translational symmetry operators. For instance, an α-helix has 3.6 residues per turn.

Besides translational and rotational symmetry operators, mirror symmetry and inversion symmetry exist. Mathematically, it can be proven that not all combinations of symmetry elements are allowed, but that 230 different combinations can occur. They are the space groups which are discussed extensively in *IT* A (2005). The graphical and printed symbols for the symmetry elements are also found in *IT* A (Chapter 1.4).

Biological macromolecules consist of building blocks such as amino acids or sugars. In general, these building-block structures are not symmetrical and the mirror images of the macromolecules do not exist in nature. Space groups with mirror planes and/or inversion centres are not allowed for crystals of these molecules, because these symmetry operations interchange right and left hands. Biological macromolecules crystallize in one of the 65 enantiomorphic space groups. (Enantiomorphic means the structure is not superimposable on its mirror image.) Apparently, some of these space groups supply more favourable packing conditions for proteins than others. The most favoured space group is $P2_12_12_1$ (Table 2.1.2.1). A consequence of symmetry is

Table 2.1.2.1
The most common space groups for protein crystals

Situation as of April 1997; data extracted from the Protein Data Bank and supplied by Rob Hooft, EMBL Heidelberg.

Space group	Occurrence (%)
$P2_12_12_1$	23
$P2_1$	11
$P3_221$	8
$P2_12_12$	6
$C2$	6

that multiple copies of particles exist in the unit cell. For instance, in space group $P2_1$ (space group No. 4), one can always expect two exactly identical entities in the unit cell, and one half of the unit cell uniquely represents the structure. This unique part of the structure is called the asymmetric unit. Of course, the asymmetric unit does not necessarily contain one protein molecule. Sometimes the unit cell contains fewer molecules than anticipated from the number of asymmetric units. This happens when the molecules occupy a position on a crystallographic axis. This is called a special position. In this situation, the molecule itself obeys the axial symmetry. Otherwise, the molecules in an asymmetric unit are on general positions. There may also be two, three or more equal or nearly equal molecules in the asymmetric unit related by noncrystallographic symmetry.

2.1.3. Point groups and crystal systems

If symmetry can be recognized in the external shape of a body, like a crystal or a virus molecule, corresponding symmetry elements have no translations, because internal translations (if they exist) do not show up in macroscopic properties. Moreover, they pass through one point, and this point is not affected by the symmetry operations (point-group symmetry). For idealized crystal shapes, the symmetry axes are limited to one-, two-, three-, four- and sixfold rotation axes because of the space-filling requirement for crystals. With the addition of mirror planes and inversion centres, there are a total of 32 possible crystallographic point groups.

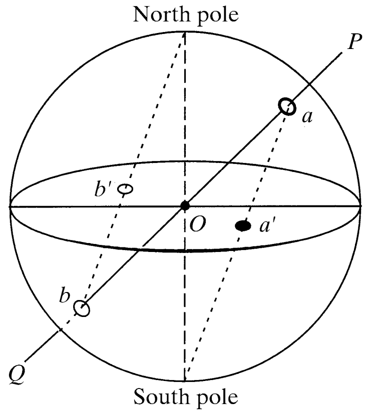


Figure 2.1.3.1
How to construct a stereographic projection. Imagine a sphere around the crystal with O as the centre. O is also the origin of the coordinate system of the crystal. Symmetry elements of the point groups pass through O. Line OP is normal to a crystal plane. It cuts through the sphere at point a. This point a is projected onto the horizontal plane through O in the following way: a vertical dashed line is drawn through O normal to the projection plane and connecting a north and a south pole. Point a is connected to the pole on the other side of the projection plane, the south pole, and is projected onto the horizontal plane at a'. For a normal OQ intersecting the lower part of the sphere, the point of intersection b is connected to the north pole and projected at b'. For the symmetry elements, their points of intersection with the sphere are projected onto the horizontal plane.

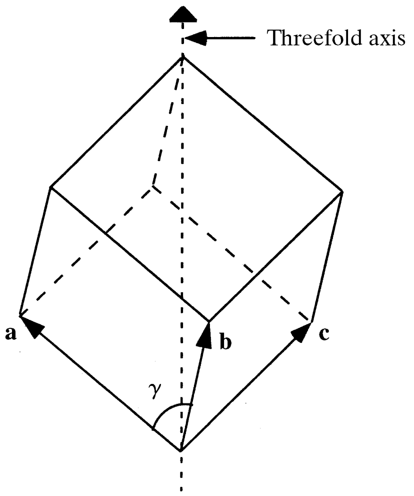


Figure 2.1.3.2
A rhombohedral unit cell.

Not all combinations of axes are allowed. For instance, a combination of two twofold axes at an arbitrary angle with respect to each other would multiply to an infinite number of twofold axes. A twofold axis can only be combined with another twofold axis at 90°. A third twofold axis is then automatically produced perpendicular to the first two (point group 222). In the same way, a threefold axis can only be combined with three twofold axes perpendicular to the threefold axis (point group 32).

For crystals of biological macromolecules, point groups with mirrors or inversion centres are not allowed, because these molecules are chiral. This restricts the number of crystallographic point groups for biological macromolecules to 11; these are the enantiomorphic point groups and are presented in Table 2.1.3.1.

Although the crystals of asymmetric molecules can only belong to one of the 11 enantiomorphic point groups, it is nevertheless important to be aware of the other point groups, especially the 11 centrosymmetric ones (Table 2.1.3.2). This is because if anomalous scattering can be neglected, the X-ray diffraction pattern of a crystal is always centrosymmetric, even if the crystal itself is asymmetric (see Sections 2.1.7 and 2.1.8).

The protein capsids of spherical virus molecules have their subunits packed in a sphere with icosahedral symmetry (532). This is the symmetry of a noncrystallographic point group (Table 2.1.3.3). A fivefold axis is allowed because translation symmetry does not apply to a virus molecule. Application of the 532 symmetry leads to 60 identical subunits in the sphere. This is the simplest type of spherical virus (triangulation number $T = 1$). Larger numbers of subunits can also be incorporated in this icosahedral surface lattice, but then the subunits lie in quasi-equivalent environments and T assumes values of 3, 4 or 7. For instance, for $T = 3$ particles there are 180 identical subunits in quasi-identical environments.

On the basis of their symmetry, the point groups are subdivided into crystal systems as follows. For each of the point groups, a set of axes can be chosen displaying the external symmetry of the crystal as clearly as possible, and, in this way, the seven crystal systems of Table 2.1.3.4 are obtained. If no other symmetry is present apart from translational symmetry, the crystal belongs to the triclinic system. With one twofold axis or screw axis, it is monoclinic. The convention in the monoclinic system is to choose the b axis along the twofold axis. The orthorhombic system

Table 2.1.3.1
The 11 enantiomorphic point groups

The point groups are presented as two stereographic projections (see Fig. 2.1.3.1). On the right is a projection of the symmetry elements, and on the left a projection of the general faces. They are arranged according to the crystal system to which they belong: triclinic, monoclinic *etc*. Different point groups are separated by full horizontal rules. The monoclinic point groups are given in two settings: in the conventional setting with the twofold axis along **b** (unique axis *b*), and the other setting with unique axis *c*. The *b* axis is horizontal in the projection plane, and the *c* axis is normal to the plane. Three-, four- and sixfold axes are always set along the *c* axis, normal to the plane. A special case is the trigonal system; either hexagonal axes or rhombohedral axes can be chosen. In the hexagonal case, the threefold axis is along the *c* axis. The other two axes are chosen along or between the twofold axes, which include an angle of 120°. In the rhombohedral setting, the threefold axis is along the body diagonal of the unit cell, and the unit cell vectors **a**, **b** and **c** are the shortest non-coplanar lattice vectors symmetrically equivalent with respect to the threefold axis (Fig. 2.1.3.2). Symbols:

● General face above the plane
○ General face below the plane
⬮ Twofold rotation axis
▲ Threefold rotation axis
■ Fourfold rotation axis
⬢ Sixfold rotation axis

Adapted with permission from *IT* A (2005), Table 10.1.2.2.

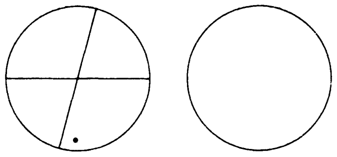
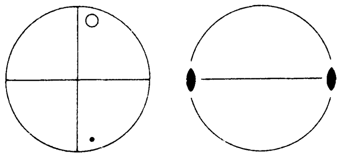
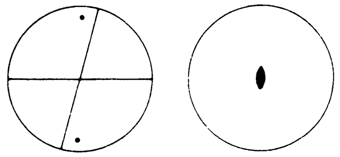
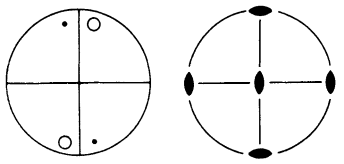
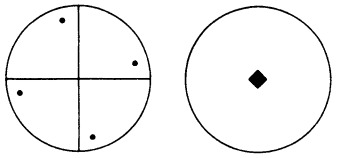
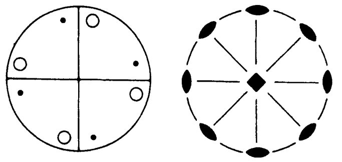
TRICLINIC	1	
MONOCLINIC	2	 Unique axis *b*  Unique axis *c*
ORTHORHOMBIC	222	
TETRAGONAL	4	
TETRAGONAL	422	

Table 2.1.3.1 (continued)

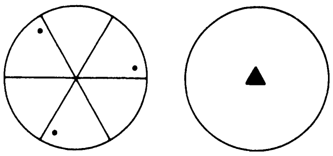
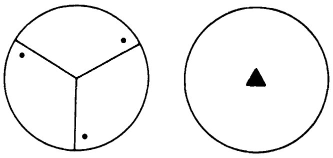
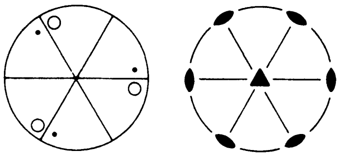
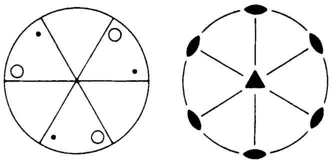
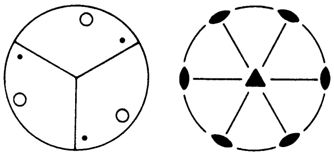
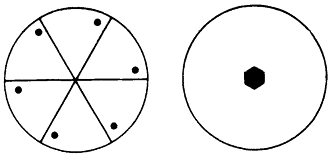
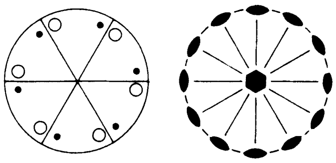
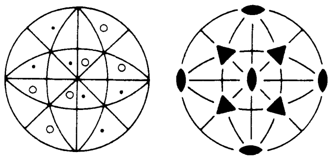
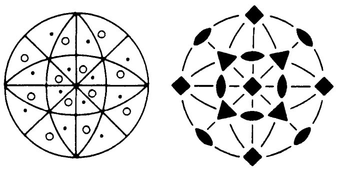
TRIGONAL Hexagonal axes	3	
TRIGONAL Rhombohedral axes	3	
TRIGONAL Hexagonal axes	321	
TRIGONAL Hexagonal axes	312	
TRIGONAL Rhombohedral axes	32	
HEXAGONAL	6	
HEXAGONAL	622	
CUBIC	23	
CUBIC	432	

Table 2.1.3.2
The 11 point groups with a centre of symmetry

For details see Table 2.1.3.1. Projections of mirror planes are indicated by a bold line or circle. The inversion centre ($\bar{1}$) is indicated by a small circle at the origin. Adapted with permission from *IT* A (2005), Table 10.1.2.2.

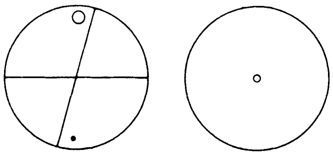
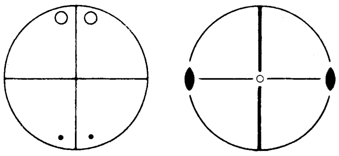
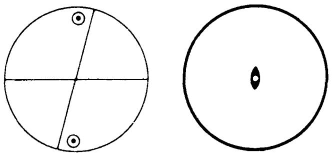
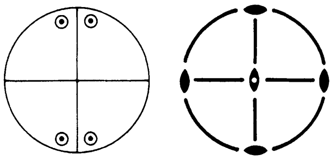
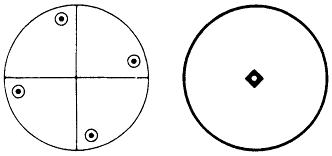
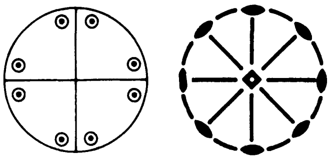
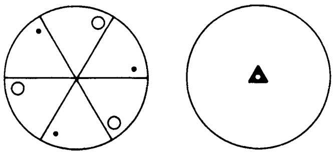
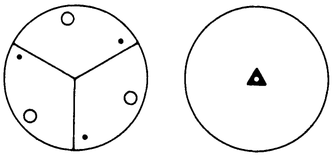
TRICLINIC	$\bar{1}$	
MONOCLINIC	$2/m$	 Unique axis *b*  Unique axis *c*
ORTHORHOMBIC	mmm or $\frac{2}{m}\frac{2}{m}\frac{2}{m}$	
TETRAGONAL	$4/m$	
TETRAGONAL	$4/mmm$ or $\frac{4}{m}\frac{2}{m}\frac{2}{m}$	
TRIGONAL Hexagonal axes	$\bar{3}$	
TRIGONAL Rhombohedral axes	$\bar{3}$	

Table 2.1.3.2 (continued)

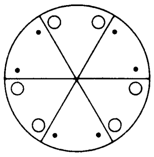
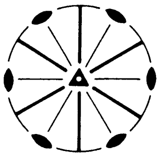
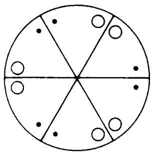
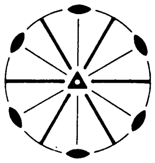
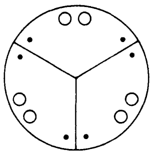
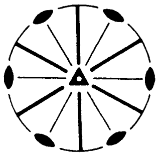
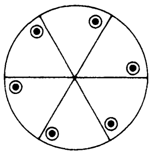
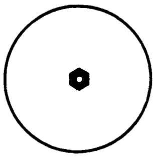
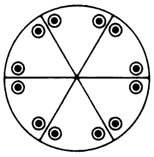
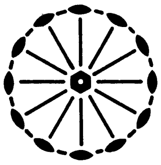
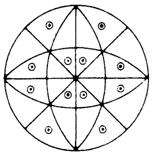
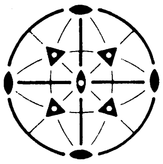
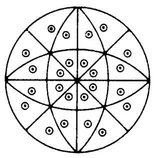
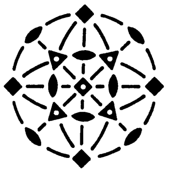
TRIGONAL Hexagonal axes	$\bar{3}m1$ or $\bar{3}\frac{2}{m}1$		
TRIGONAL Hexagonal axes	$\bar{3}1m$ or $\bar{3}1\frac{2}{m}$		
TRIGONAL Rhombohedral axes	$\bar{3}m$ or $\bar{3}\frac{2}{m}$		
HEXAGONAL	$6/m$		
HEXAGONAL	$6/mmm$ or $\frac{6}{m}\frac{2}{m}\frac{2}{m}$		
CUBIC	$m\bar{3}$ or $\frac{2}{m}\bar{3}$		
CUBIC	$m\bar{3}m$ or $\frac{4}{m}\bar{3}\frac{2}{m}$		

Table 2.1.3.3
The icosahedral point group 532

For details see Table 2.1.3.1. Adapted with permission from *IT* A (2005), Table 10.1.4.3.

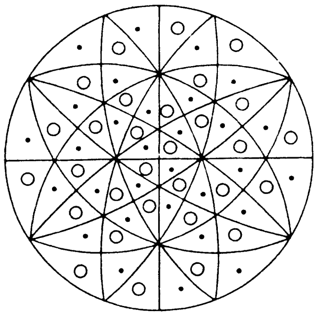
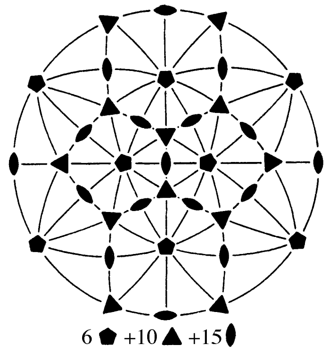
ICOSAHEDRAL	532		 6⬟ +10▲ +15⬮

Table 2.1.3.4
The seven crystal systems

Crystal system	Conditions imposed on cell geometry	Minimum point-group symmetry
Triclinic	None	1
Monoclinic	Unique axis b: $\alpha = \gamma = 90°$	2
Orthorhombic	$\alpha = \beta = \gamma = 90°$	222
Tetragonal	$a = b$; $\alpha = \beta = \gamma = 90°$	4
Trigonal	Hexagonal axes: $a = b$; $\alpha = \beta = 90°$; $\gamma = 120°$ Rhombohedral axes: $a = b = c$; $\alpha = \beta = \gamma$ †	3
Hexagonal	$a = b$; $\alpha = \beta = 90°$; $\gamma = 120°$	6
Cubic	$a = b = c$; $\alpha = \beta = \gamma = 90°$	23

† A rhombohedral unit cell can be regarded as a cube extended or compressed along the body diagonal (the threefold axis) (see Fig. 2.1.3.2).

has three mutually perpendicular twofold (screw) axes. Another convention is that in tetragonal, trigonal and hexagonal crystals, the axis of highest symmetry is labelled c. These conventions can deviate from the guide rules for unit-cell choice given in Section 2.1.1.

The seven crystal systems are based on the point-group symmetry. Except for the triclinic unit cell, all other cells can occur either as primitive unit cells or as centred unit cells (Section 2.1.1). A total of 14 different types of unit cell exist, depicted in Fig. 2.1.3.3. Their corresponding crystal lattices are commonly called Bravais lattices.

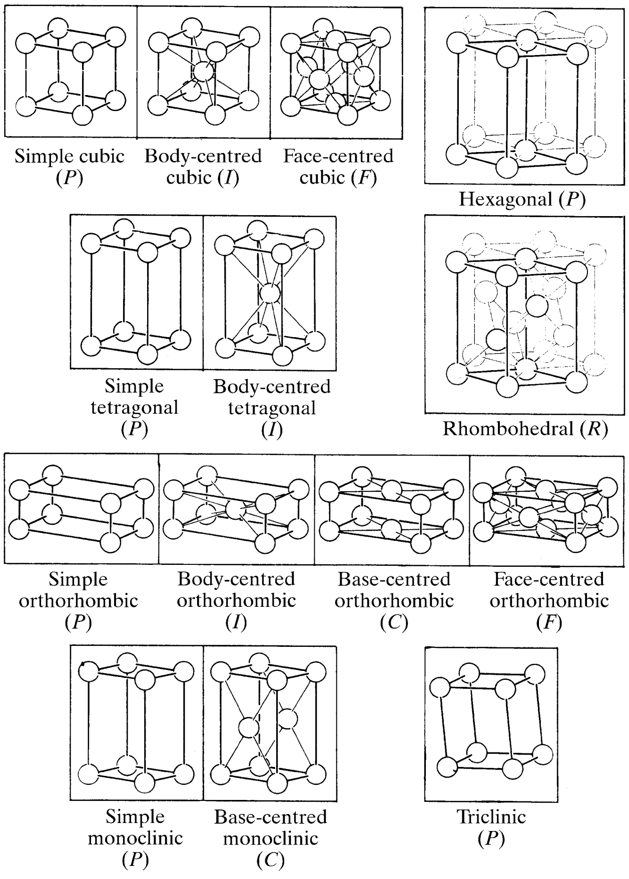

Simple cubic (*P*) Body-centred cubic (*I*) Face-centred cubic (*F*) Hexagonal (*P*)
Simple tetragonal (*P*) Body-centred tetragonal (*I*) Rhombohedral (*R*)
Simple orthorhombic (*P*) Body-centred orthorhombic (*I*) Base-centred orthorhombic (*C*) Face-centred orthorhombic (*F*)
Simple monoclinic (*P*) Base-centred monoclinic (*C*) Triclinic (*P*)

Figure 2.1.3.3
The 14 Bravais lattices. Reproduced with permission from Burzlaff & Zimmermann (2005).

2.1.4. Basic diffraction physics

2.1.4.1. Diffraction by one electron

The scattering of an X-ray beam by a crystal results from interaction between the electric component of the beam and the electrons in the crystal. The magnetic component has hardly any effect and can be disregarded.

If a monochromatic polarized beam hits an electron, the electron starts to oscillate in the direction of the electric vector of the incident beam (Fig. 2.1.4.1). This oscillating electron acts as the aerial of a transmitter and radiates X-rays with the same or lower frequency as the incident beam. The frequency change is due to the Compton effect: the photons of the incident beam collide with the electron and lose part of their energy. This is inelastic scattering, and the scattered radiation is incoherent with the incident beam. Compton scattering contributes to the background in a diffraction experiment. In elastic scattering, the scattered radiation has the same wavelength as the incident

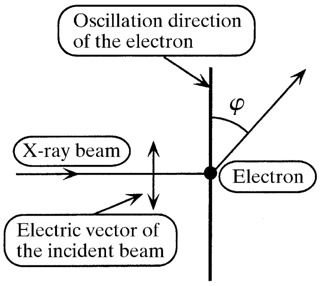

Oscillation direction of the electron
φ
X-ray beam
Electron
Electric vector of the incident beam

Figure 2.1.4.1
The electric vector of a monochromatic and polarized X-ray beam is in the plane. It hits an electron, which starts to oscillate in the same direction as the electric vector of the beam. The oscillating electron acts as a source of X-rays. The scattered intensity depends on the angle φ between the oscillation direction of the electron and the scattering direction [equation (2.1.4.1)]. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

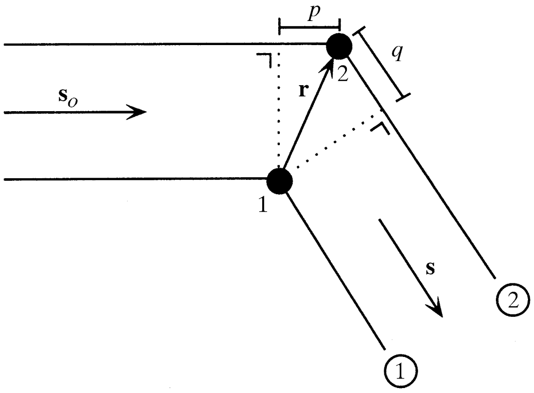


Figure 2.1.4.2
The black dots are electrons. The origin of the system is at electron 1; electron 2 is at position **r**. The electrons are irradiated by an X-ray beam from the direction indicated by vector $\mathbf{s}_o$. The radiation scattered by the electrons is observed in the direction of vector **s**. Because of the path difference $p+q$, scattered beam 2 will lag behind scattered beam 1 in phase. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

radiation, and this is the radiation responsible for the interference effects in diffraction. It was shown by Thomson that if the electron is completely free the following hold:

(1) The phase difference between the incident and the scattered beam is π, because the scattered radiation is proportional to the displacement of the electron, which differs by π in phase with its acceleration imposed by the electric vector.
(2) The amplitude of the electric component of the scattered wave at a distance r which is large in comparison with the wavelength of the radiation is

$$E_{\rm el} = E_o \frac{1}{r}\frac{e^2}{mc^2}\sin\varphi,$$

where E_o is the amplitude of the electric vector of the incident beam, e is the electron charge, m is its mass, c is the speed of light and φ is the angle between the oscillation direction of the electron and the scattering direction (Fig. 2.1.4.1). Note that $E_o \sin\varphi$ is the component of E_o perpendicular to the scattering direction.

In terms of energy,

$$I_{\rm el} = I_o \frac{1}{r^2}\left(\frac{e^2}{mc^2}\right)^2 \sin^2\varphi. \qquad (2.1.4.1a)$$

The scattered energy per unit solid angle is

$$I_{\rm el}(\Omega = 1) = I_{\rm el} r^2. \qquad (2.1.4.1b)$$

It was shown by Klein & Nishina (1929) [see also Heitler (1966)] that the scattering by an electron can be discussed in terms of the classical Thomson scattering if the quantum energy $h\nu \ll mc^2$. This is not true for very short X-ray wavelengths. For $\lambda = 0.0243$ Å, $h\nu$ and mc^2 are exactly equal, but for $\lambda = 1.0$ Å, $h\nu$ is 0.0243 times mc^2. Since wavelengths in macromolecular crystallography are usually in the range 0.8–2.5 Å, the classical approximation is allowed. It should be noted that:

(1) The intensity scattered by a free electron is independent of the wavelength.
(2) Thomson's equation can also be applied to other charged particles, *e.g.* a proton. Because the mass of a proton is 1800 times the electron mass, scattering by protons and by atomic nuclei can be neglected.
(3) Equation (2.1.4.1*a*) gives the scattering for a polarized beam. For an unpolarized beam, $\sin^2\varphi$ is replaced by a suitable polarization factor.

2.1.4.2. *Scattering by a system of two electrons*

This can be derived along classical lines by calculating the phase difference between the X-ray beams scattered by each of the two electrons. A derivation based on quantum mechanics leads exactly to the same result by calculating the transition probability for the scattering of a primary quantum $(h\nu)_o$, given a secondary quantum $h\nu$ (Heitler, 1966, p. 193). For simplification we shall give only the classical derivation here. In Fig. 2.1.4.2, a system of two electrons is drawn with the origin at electron 1 and electron 2 at position **r**. They scatter the incident beam in a direction given by the vector **s**. The direction of the incident beam is along the vector $\mathbf{s}_o$. The length of the vectors can be chosen arbitrarily, but for convenience they are given a length $1/\lambda$. The two electrons scatter completely independently of each other.

Therefore, the amplitudes of the scattered beams 1 and 2 are equal, but they have a phase difference resulting from the path difference between the beam passing through electron 2 and the beam passing through electron 1. The path difference is $p + q = \lambda[\mathbf{r}\cdot(\mathbf{s}_o - \mathbf{s})]$. Beam 2 lags behind in phase compared with beam 1, and with respect to wave 1 its phase angle is

$$-2\pi\lambda[\mathbf{r}\cdot(\mathbf{s}_o - \mathbf{s})]/\lambda = 2\pi\mathbf{r}\cdot\mathbf{S}, \qquad (2.1.4.2)$$

where $\mathbf{S} = \mathbf{s} - \mathbf{s}_o$.

From Fig. 2.1.4.3, it is clear that the direction of **S** is perpendicular to an imaginary plane reflecting the incident beam at an angle θ and that the length of **S** is given by

$$|\mathbf{S}| = 2\sin\theta/\lambda. \qquad (2.1.4.3)$$

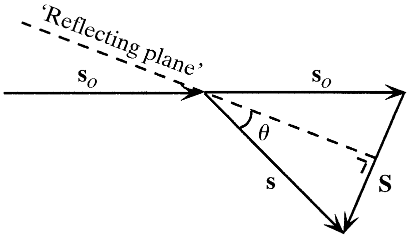


Figure 2.1.4.3
The direction of the incident wave is indicated by $\mathbf{s}_o$ and that of the scattered wave by **s**. Both vectors are of length $1/\lambda$. A plane that makes equal angles with **s** and $\mathbf{s}_o$ can be regarded as a mirror reflecting the incident beam. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

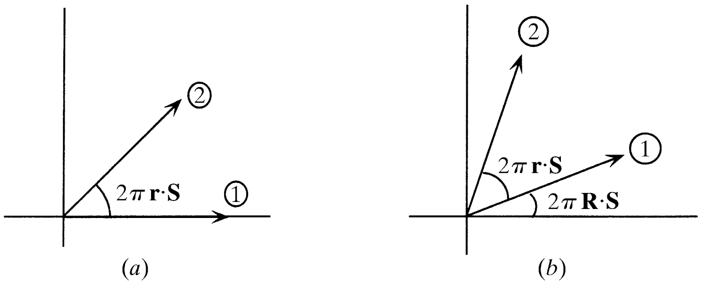

②
$2\pi\mathbf{r}\cdot\mathbf{S}$
①
(a)
②
$2\pi\mathbf{r}\cdot\mathbf{S}$
①
$2\pi\mathbf{R}\cdot\mathbf{S}$
(b)

Figure 2.1.4.4
An Argand diagram for the scattering by two electrons. In (*a*), the origin is at electron 1; electron 2 is at position **r** with respect to electron 1. In (*b*), electron 1 is at position **R** with respect to the new origin, and electron 2 is at position $\mathbf{R} + \mathbf{r}$.

The total scattering from the two-electron system is $1 + 1 \times \exp(2\pi i\mathbf{r}\cdot\mathbf{S})$ if the resultant amplitude of the waves from electrons 1 and 2 is set to 1. In an Argand diagram, the waves are represented by vectors in a two-dimensional plane, as in Fig. 2.1.4.4(*a*).[1] Thus far, the origin of the system was chosen at electron 1. Moving the origin to another position simply means an equal change of phase angle for all waves. Neither the amplitudes nor the intensities of the reflected beams change (Fig. 2.1.4.4*b*).

2.1.4.3. *Scattering by atoms*

2.1.4.3.1. *Scattering by one atom*

Electrons in an atom are bound by the nucleus and are – in principle – not free electrons.

However, to a good approximation, they can be regarded as such if the frequency of the incident radiation ν is greater than the natural absorption frequencies, ν_n, at the absorption edges of the scattering atom, or the wavelength of the incident radiation is shorter than the absorption-edge wavelength (Section 2.1.4.4). This is normally true for light atoms but not for heavy ones (Table 2.1.4.1).

Table 2.1.4.1
The position of the $K\alpha$ edge of different elements

Atomic number	Element	$K\alpha$ edge (Å)
6	C	43.68
16	S	5.018
26	Fe	1.743
34	Se	0.980
78	Pt	0.158

If the electrons in an atom can be regarded as free electrons, the scattering amplitude of the atom is a real quantity, because the electron cloud has a centrosymmetric distribution, *i.e.* $\rho(\mathbf{r}) = \rho(-\mathbf{r})$.

A small volume, $\mathrm{d}v_r$, at **r** contains $\rho(\mathbf{r}) \times \mathrm{d}v_r$ electrons, and at $-\mathbf{r}$ there are $\rho(-\mathbf{r}) \times \mathrm{d}v_r$ electrons. The combined scattering of the two volume elements, in units of the scattering of a free electron, is

$$\rho(\mathbf{r})\mathrm{d}v_r\{\exp(2\pi i\mathbf{r}\cdot\mathbf{S}) + \exp[2\pi i(-\mathbf{r})\cdot\mathbf{S}]\} = 2\rho(\mathbf{r})\cos(2\pi\mathbf{r}\cdot\mathbf{S})\mathrm{d}v_r;$$

this is a real quantity.

The scattering amplitude of an atom is called the atomic scattering factor *f*. It expresses the scattering of an atom in terms of the scattering of a single electron. *f* values are calculated for spherically averaged electron-density distributions and, therefore, do not depend on the scattering direction. They are tabulated in *IT* C (2004) as a function of $\sin\theta/\lambda$. The *f* values decrease appreciably as a function of $\sin\theta/\lambda$ (Fig. 2.1.4.5). This is due to interference effects between the scattering from the electrons in the cloud. In the direction $\theta = 0$, all electrons scatter in phase and the atomic scattering factor is equal to the number of electrons in the atom.

2.1.4.3.2. *Scattering by a plane of atoms*

A plane of atoms reflects an X-ray beam with a phase retardation of $\pi/2$ with respect to the scattering by a single atom. The difference is caused by the difference in path length from source (*S*) to atom (*M*) to detector (*D*) for the different atoms in the plane (Fig. 2.1.4.6). Suppose the plane is infinitely large. The shortest connection between *S* and *D via* the plane is *S–M–D*. The plane containing *S*, *M* and *D* is perpendicular to the reflecting plane, and the lines *SM* and *MD* form equal angles with the reflecting plane. Moving outwards from atom *M* in the reflecting plane, to *P* for instance, the path length *S–P–D* is longer. At the edge of the first Fresnel zone, the path is $\lambda/2$ longer (Fig. 2.1.4.6). This edge is an ellipse with its centre at *M* and its major axis on the line of intersection between the plane *SMD* and the reflecting plane. Continuing outwards, many more elliptic Fresnel zones are formed. Clearly, the beams radiated by the many atoms in the plane interfere with each other. The situation is represented in the Argand diagram in Fig. 2.1.4.7. Successive Fresnel zones can be subdivided into an equal number of subzones. If the distribution of electrons is sufficiently homogeneous, it can be assumed that the subzones in one Fresnel zone give the same amplitude at *D*. Their phases are spaced at regular intervals and their vectors in the Argand diagram lie in a half circle. In the lower part of Fig. 2.1.4.7, this is illustrated for the first Fresnel zone. For the second Fresnel zone (upper part),

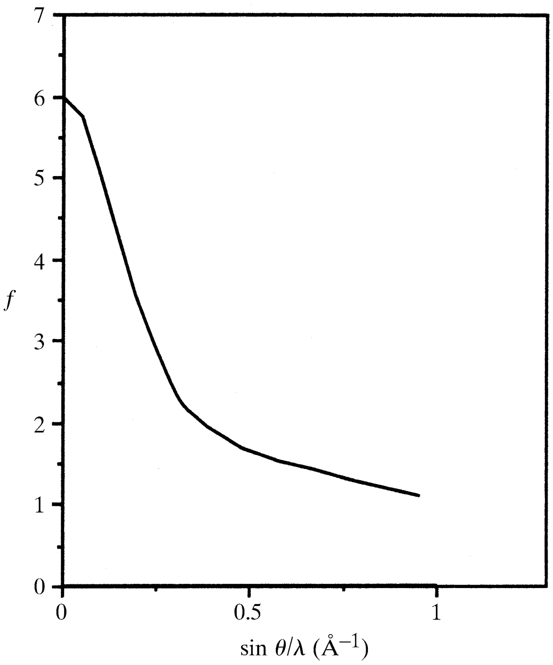


Figure 2.1.4.5
The atomic scattering factor *f* for carbon as a function of $\sin\theta/\lambda$, expressed in units of the scattering by one electron. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

---

[1] The plane is also called the 'imaginary plane'. The real part of the vector in an Argand diagram is along the horizontal or real axis; the imaginary part is along the vertical or imaginary axis. Note also that $\exp(2\pi i\mathbf{r}\cdot\mathbf{S}) = \cos(2\pi\mathbf{r}\cdot\mathbf{S}) + i\sin(2\pi\mathbf{r}\cdot\mathbf{S})$. The cosine term is the real component and the sine term is the imaginary component.

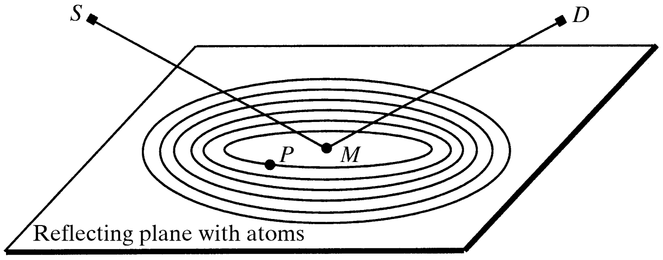


Figure 2.1.4.6
S is the X-ray source and D is the detector. The scattering is by the atoms in a plane. The shortest distance between S and D *via* a point in the plane is through M. Path lengths *via* points in the plane further out from M are longer, and when these beams reach the detector they lag behind in phase with respect to the MD beam. The plane is divided into zones, such that from one zone to the next the path difference is $\lambda/2$.

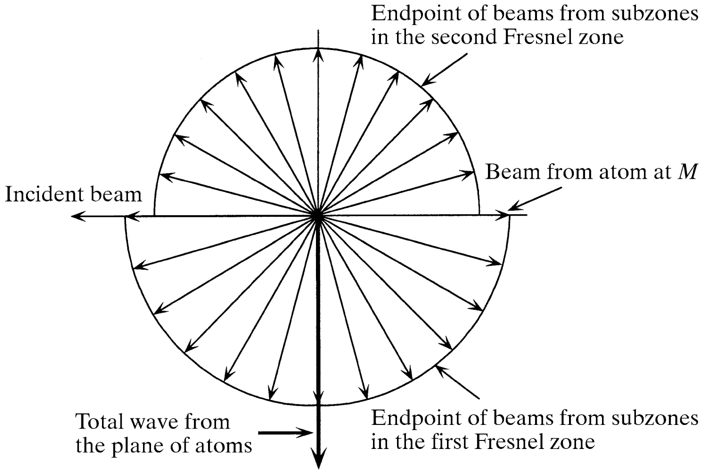

Endpoint of beams from subzones in the second Fresnel zone
Beam from atom at M
Incident beam
Total wave from the plane of atoms
Endpoint of beams from subzones in the first Fresnel zone

Figure 2.1.4.7
Schematic picture of the Argand diagram for the scattering by atoms in a plane. All electrons are considered free. The vector of the incident beam points to the left. The atom at M (see Fig. 2.1.4.6) has a phase difference of π with respect to the incident beam. Subzones in the first Fresnel zone have the endpoints of their vectors on the lower half circle. For the next Fresnel zone, they are on the upper half circle, which has a smaller radius because the amplitude decreases gradually for subsequent Fresnel zones (Kauzmann, 1957). The sum of all vectors points down, indicating a phase lag of $\pi/2$ with respect to the beam scattered by the atom at M.

the radius is slightly smaller, because the intensity radiated by more distant zones decreases (Kauzmann, 1957). Therefore, the sum of vectors pointing upwards is shorter than that of those pointing downwards, and the resulting scattered wave lags $\pi/2$ in phase behind the scattering by the atom at M.

2.1.4.4. *Anomalous dispersion*

In classical dispersion theory, the scattering power of an atom is derived by supposing that the atom contains dipole oscillators. In units of the scattering of a free electron, the scattering of an oscillator with eigen frequency ν_n and moderate damping factor κ_n was found to be a complex quantity:

$$f_n = \nu^2/(\nu^2 - \nu_n^2 - i\kappa_n\nu), \qquad (2.1.4.4)$$

where ν is the frequency of the incident radiation [James, 1965; see also *IT* C (2004), equation (4.2.6.8)]. When $\nu \gg \nu_n$ in equation (2.1.4.4), f_n approaches unity, as is the case for scattering by a free electron; when $\nu \ll \nu_n$, f_n approaches zero, demonstrating the lack of scattering from a fixed electron. Only for $\nu \cong \nu_n$ does the imaginary part have an appreciable value.

Fortunately, quantum mechanics arrives at the same result by adding a rational meaning to the damping factors and interpreting ν_n as absorption frequencies of the atom (Hönl, 1933). For heavy atoms, the most important transitions are to a continuum of energy states, with $\nu_n \geq \nu_K$ or $\nu_n \geq \nu_L$ *etc.*, where ν_K and ν_L are the frequencies of the K and L absorption edges.

In practice, the complex atomic scattering factor, $f_{\rm anomalous}$, is separated into three parts: $f_{\rm anomalous} = f + f' + if''$. f is the contribution to the scattering if the electrons are free electrons and it is a real number (Section 2.1.4.3). f' is the real part of the correction to be applied and f'' is the imaginary correction; f'' is always $\pi/2$ in phase ahead of f (Fig. 2.1.4.8). $f + f'$ is the total real part of the atomic scattering factor.

The imaginary correction if'' is connected with absorption by oscillators having $\nu_n \cong \nu$. It can be calculated from the atomic absorption coefficient of the anomalously scattering element. For each of the K, L *etc.* absorption edges, f'' is virtually zero for frequencies below the edge, but it rises steeply at the edge and decreases gradually at higher frequencies.

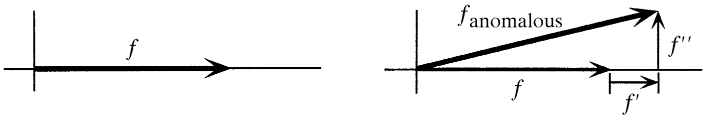
The real correction f' can be derived from f'' by means of the Kramers–Kronig transform [*IT* C (2004), Section 4.2.6.2.2]. For frequencies close to an absorption edge, f' becomes strongly negative.

Values for f, f' and f'' are always given in units equal to the scattering by one free electron. f values are tabulated in *IT* C (2004) as a function of $\sin\theta/\lambda$, and the anomalous-scattering corrections for forward scattering as a function of the wavelength. Because the anomalous contribution to the atomic scattering factor is mainly due to the electrons close to the nucleus, the value of the corrections diminishes much more slowly than f as a function of the scattering angle.

2.1.4.5. *Scattering by a crystal*

A unit cell contains a large number of electrons, especially in the case of biological macromolecules. The waves scattered by these electrons interfere with each other, thereby reducing the effective number of electrons in the scattered wave. The exception is scattering in the forward direction, where the beams from all electrons are in phase and add to each other. The effective number of scattering electrons is called the *structure factor F* because it depends on the structure, *i.e.* the distribution of the atoms in the unit cell. It also depends on the scattering direction. If small electron-density changes due to chemical bonding are neglected, the structure factor can be regarded as the sum of the scattering by the atoms in the unit cell, taking into consideration their positions and the corresponding phase differences between the scattered waves. For n atoms in the unit cell

$$F(\mathbf{S}) = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{r}_j \cdot \mathbf{S}), \qquad (2.1.4.5)$$



No anomalous scattering
(*a*)

Anomalous scattering
(*b*)

Figure 2.1.4.8
The atomic scattering factor as a vector in the Argand diagram. (*a*) When the electrons in the atom can be regarded as free. (*b*) When they are not completely free and the scattering becomes anomalous with a real anomalous contribution f' and an imaginary contribution if''. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

where **S** is a vector perpendicular to the plane reflecting the incident beam at an angle θ; the length of **S** is given by $|\mathbf{S}| = 2\sin\theta/\lambda$ [equation (2.1.4.3) in Section 2.1.4.2].

The origin of the system is chosen at the origin of the selected unit cell. Atom j is at position $\mathbf{r}_j$ with respect to the origin. Another unit cell has its origin at $t \times \mathbf{a}$, $u \times \mathbf{b}$ and $v \times \mathbf{c}$, where t, u and v are whole numbers, and **a**, **b** and **c** are the basis vectors of the unit cell. With respect to the first origin, its scattering is

$$F(\mathbf{S}) \exp(2\pi i t\mathbf{a} \cdot \mathbf{S}) \exp(2\pi i u\mathbf{b} \cdot \mathbf{S}) \exp(2\pi i v\mathbf{c} \cdot \mathbf{S}).$$

The wave scattered by a crystal is the sum of the waves scattered by all unit cells. Assuming that the crystal has a very large number of unit cells ($n_1 \times n_2 \times n_3$), the amplitude of the wave scattered by the crystal is

$$W_{\rm cr}(\mathbf{S}) = F(\mathbf{S}) \sum_{t=0}^{n_1} \exp(2\pi i t\mathbf{a} \cdot \mathbf{S}) \sum_{u=0}^{n_2} \exp(2\pi i u\mathbf{b} \cdot \mathbf{S}) \times \sum_{v=0}^{n_3} \exp(2\pi i v\mathbf{c} \cdot \mathbf{S}). \qquad (2.1.4.6)$$

For an infinitely large crystal, the three summations over the exponential functions are delta functions. They have the property that they are zero unless

$$\mathbf{a} \cdot \mathbf{S} = h, \ \mathbf{b} \cdot \mathbf{S} = k \text{ and } \mathbf{c} \cdot \mathbf{S} = l, \qquad (2.1.4.7)$$

where h, k and l are whole numbers, either positive, negative or zero. These are the Laue conditions. If they are fulfilled, all unit cells scatter in phase and *the amplitude of the wave scattered by the crystal is proportional to the amplitude of the structure factor F*. Its intensity is proportional to $|F|^2$.

S vectors satisfying equation (2.1.4.7) are denoted by **S**(*hkl*) or **S**(**h**), and the corresponding structure factors as $F(hkl)$ or $F(\mathbf{h})$.

Bragg's law for scattering by a crystal is better known than the Laue conditions:

$$2d \sin\theta = \lambda, \qquad (2.1.4.8)$$

where d is the distance between reflecting lattice planes, θ is the reflecting or glancing angle and λ is the wavelength (Fig. 2.1.4.9). It can easily be shown that the Laue conditions and Bragg's law are equivalent by combining equation (2.1.4.7) with the following information:

(1) Vector **S** is perpendicular to a reflecting plane (Section 2.1.4.2).

(2) The Laue conditions for scattering [equation (2.1.4.7)] can be written as

$$\frac{\mathbf{a}}{h} \cdot \mathbf{S} = 1; \quad \frac{\mathbf{b}}{k} \cdot \mathbf{S} = 1; \quad \frac{\mathbf{c}}{l} \cdot \mathbf{S} = 1. \qquad (2.1.4.9)$$

(3) Lattice planes always divide the unit-cell vectors **a**, **b** and **c** into a number of equal parts (Section 2.1.1). If the lattice planes divide the **a** vector of the unit cell into h equal parts, the first index for this set of planes is h. The second index, k, is related to the division of **b** and the third index, l, to the division of **c**.

From equation (2.1.4.9) it follows that vector **S**(*hkl*) is perpendicular to a plane determined by the points **a**/*h*, **b**/*k* and **c**/*l*, and according to conditions (3) this is a lattice plane. Therefore, scattering by a crystal can indeed be regarded as reflection by lattice planes. The projection of **a**/*h*, **b**/*k* and **c**/*l* on vector **S**(*hkl*) is $1/|\mathbf{S}(hkl)|$ (Laue condition), but it is also equal to the spacing $d(hkl)$ between the lattice planes (see Fig. 2.1.1.3), and, therefore, $|\mathbf{S}(hkl)| = 1/d(hkl)$. Combining this with equation (2.1.4.3) yields Bragg's law, $2d\sin\theta = \lambda$ [equation (2.1.4.8)].

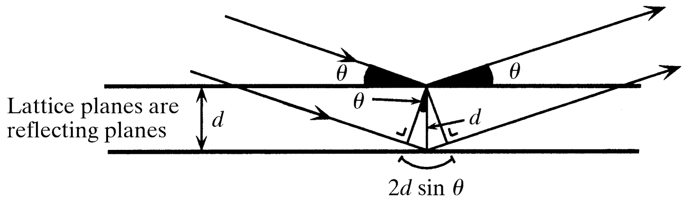


Figure 2.1.4.9
X-ray diffraction by a crystal is, in Bragg's conception, reflection by lattice planes. The beams reflected by successive planes have a path difference of $2d\sin\theta$, where d is the lattice-plane distance and θ is the reflecting angle. Positive interference occurs if $2d\sin\theta = \lambda, 2\lambda, 3\lambda$ *etc.*, where λ is the X-ray wavelength. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

2.1.4.6. *The structure factor*

For noncentrosymmetric structures, the structure factor,

$$F(\mathbf{S}) = \sum_{j=1}^{n} f_j \exp(2\pi i \mathbf{r}_j \cdot \mathbf{S}),$$

is an imaginary quantity and can also be written as[2]

$$F(\mathbf{S}) = \sum_{j=1}^{n} f_j \cos(2\pi \mathbf{r}_j \cdot \mathbf{S}) + i \sum_{j=1}^{n} f_j \sin(2\pi \mathbf{r}_j \cdot \mathbf{S}) = A(\mathbf{S}) + iB(\mathbf{S}).$$

It is sometimes convenient to split the structure factor into its real part, $A(\mathbf{S})$, and its imaginary part, $B(\mathbf{S})$. For centrosymmetric structures, $B(\mathbf{S}) = 0$ if the origin of the structure is chosen at the centre of symmetry.

The average value of the structure-factor amplitude $|F(\mathbf{S})|$ decreases with increasing $|\mathbf{S}|$ or, because $|\mathbf{S}| = 2\sin\theta/\lambda$, with increasing reflecting angle θ.

This is caused by two factors:

(1) A stronger negative interference between the electrons in the atoms at a larger scattering angle; this is expressed in the decrease of the atomic scattering factor as a function of **S**.

(2) The temperature-dependent vibrations of the atoms. Because of these vibrations, the apparent size of an atom is larger during an X-ray exposure, and the decrease in its scattering as a function of **S** is stronger. If the vibration is equally strong in all directions, it is called isotropic, and the atomic scattering factor must be multiplied by a correction factor, the temperature factor, $\exp[-B(\sin^2\theta)/\lambda^2]$. It can be shown that the parameter B is related to the mean-square displacement of the atomic vibrations, $\overline{u^2}$:

$$B = 8\pi^2\overline{u^2}.$$

In protein crystal structures determined at high resolution, each atom is given its own individual thermal parameter B.[3] Anisotropic thermal vibration is described by six parameters instead of one, and the evaluation of this anisotropic thermal vibration requires more data (X-ray intensities) than are usually available. Only at very high resolution (better than 1.5 Å) can one consider the incorporation of anisotropic temperature factors.

---

[2] For convenience, we write $F(\mathbf{S})$ when we mean $F(hkl)$ or $F(\mathbf{h})$, and **S** instead of **S**(*hkl*) or **S**(**h**).

[3] Do not confuse the thermal parameter B with the imaginary part $B(\mathbf{S})$ of the structure factor.

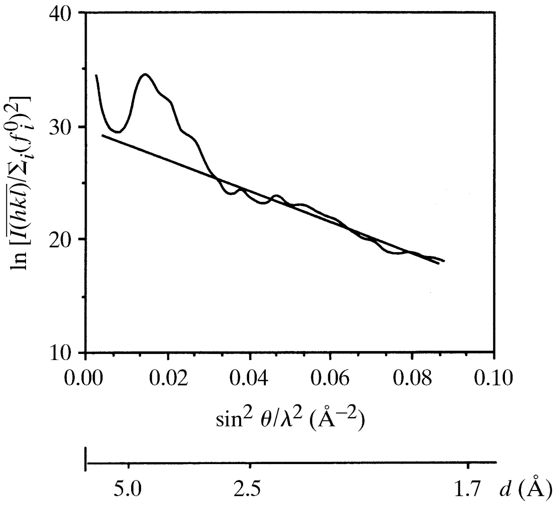


Figure 2.1.4.10
The Wilson plot for phospholipase A_2 with data to 1.7 Å resolution. Only beyond 3 Å resolution is it possible to fit the curve to a straight line. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

The value of $|F(\mathbf{S})|$ can be regarded as the effective number of electrons per unit cell scattering in the direction corresponding to $\mathbf{S}$. This is true if the values of $|F(\mathbf{S})|$ are on an absolute scale; this means that the unit of scattering is the scattering by one electron in a specific direction. The experimental values of $|F(\mathbf{S})|$ are normally on an arbitrary scale. The average value of the scattered intensity, $\overline{I(\text{abs.}, \mathbf{S})}$, on an absolute scale is $\overline{I(\text{abs.}, \mathbf{S})} = \overline{|F(\mathbf{S})|^2} = \sum_i f_i^2$, where f_i is the atomic scattering factor reduced by the temperature factor. This can be understood as follows:

$$\begin{aligned} I(\text{abs.}, \mathbf{S}) &= F(\mathbf{S}) \cdot F^*(\mathbf{S}) = |F(\mathbf{S})|^2 \\ &= \sum_i \sum_j f_i f_j \exp\big[2\pi i(\mathbf{r}_i - \mathbf{r}_j) \cdot \mathbf{S}\big]. \end{aligned} \qquad (2.1.4.10)$$

For a large number of reflections, $\mathbf{S}$ varies considerably, and assuming that the angles $[2\pi(\mathbf{r}_i - \mathbf{r}_j) \cdot \mathbf{S}]$ are evenly distributed over the range 0–2π for $i \neq j$, the average value for the terms with $i \neq j$ will be zero and only the terms with $i = j$ remain, giving

$$\overline{|F(\mathbf{S})|^2} = \overline{I(\text{abs.}, \mathbf{S})} = \sum_i f_i^2. \qquad (2.1.4.11)$$

Because of the thermal vibrations

$$f_i^2 = \exp\big(-2B_i \sin^2\theta/\lambda^2\big)(f_i^o)^2,$$

where i denotes a specific atom and f_i^o is the scattering factor for the atom i at rest.

It is sometimes necessary to transform the intensities and the structure factors from an arbitrary to an absolute scale. Wilson (1942) proposed a method for estimating the required scale factor K and, as an additional bonus, the thermal parameter B averaged over the atoms:

$$\overline{I(\mathbf{S})} = K\overline{I(\text{abs.}, \mathbf{S})} = K \exp(-2B \sin^2\theta/\lambda^2) \sum_i (f_i^o)^2. \qquad (2.1.4.12)$$

To determine K and B, equation (2.1.4.11) is written in the form

$$\ln[\overline{I(\mathbf{S})}/ \textstyle\sum_i (f_i^o)^2] = \ln K - 2B \sin^2\theta/\lambda^2. \qquad (2.1.4.13)$$

Because f_i^o depends on $\sin\theta/\lambda$, average intensities, $\overline{I(\mathbf{S})}$, are calculated for shells of narrow $\sin\theta/\lambda$ ranges. $\ln[\overline{I(\mathbf{S})}/ \sum_i (f_i^o)^2]$ is plotted against $\sin^2\theta/\lambda^2$. The result should be a straight line with slope $-2B$, intersecting the vertical axis at ln K (Fig. 2.1.4.10).

For proteins, the Wilson plot gives rather poor results because the assumption in deriving equation (2.1.4.11) that the angles, $[2\pi(\mathbf{r}_i - \mathbf{r}_j) \cdot \mathbf{S}]$, are evenly distributed over the range 0–2π for $i \neq j$ is not quite valid, especially not in the $\sin\theta/\lambda$ ranges at low resolution.

As discussed above, the average value of the structure factors, $F(\mathbf{S})$, decreases with the scattering angle because of two effects:

(1) the decrease in the atomic scattering factor f;
(2) the temperature factor.

This decrease is disturbing for statistical studies of structure-factor amplitudes. It is then an advantage to eliminate these effects by working with normalized structure factors, $E(\mathbf{S})$, defined by

$$\begin{aligned} E(\mathbf{S}) &= F(\mathbf{S}) \Big/ \Big(\sum_j f_j^2\Big)^{1/2} \\ &= F(\mathbf{S}) \exp\big(B \sin^2\theta/\lambda^2\big) \Big/ \Big[\sum_j (f_j^o)^2\Big]^{1/2}. \end{aligned} \qquad (2.1.4.14)$$

The application of equation (2.1.4.14) to $\overline{|E(\mathbf{S})|^2}$ gives

$$\overline{|E(\mathbf{S})|^2} = \overline{|F(\mathbf{S})|^2} \Big/ \sum_j f_j^2 = \overline{|F(\mathbf{S})|^2} \Big/ \overline{|F(\mathbf{S})|^2} = 1. \qquad (2.1.4.15)$$

The average value, $\overline{|E(\mathbf{S})|^2}$, is equal to 1. The advantage of working with normalized structure factors is that the scaling is not important, because if equation (2.1.4.14) is written as

$$E(\mathbf{S}) = \frac{F(\mathbf{S})}{(\overline{|F(\mathbf{S})|^2})^{1/2}},$$

a scale factor affects numerator and denominator equally.

In practice, the normalized structure factors are derived from the observed data as follows:

$$E(\mathbf{S}) = F(\mathbf{S}) \exp\big(B \sin^2\theta/\lambda^2\big) \big/ \big(\varepsilon \overline{|F(\mathbf{S})|^2}\big)^{1/2}, \qquad (2.1.4.16)$$

where ε is a correction factor for space-group symmetry. For general reflections it is 1, but it is greater than 1 for reflections having $\mathbf{h}$ parallel to a symmetry element. This can be understood as follows. For example, if m atoms are related by this symmetry element, $\mathbf{r}_j \cdot \mathbf{S}$ (with j from 1 to m) is the same in their contribution to the structure factor

$$F(\mathbf{h}) = \sum_{j=1}^{m} f_j \exp(2\pi i \mathbf{r}_j \cdot \mathbf{S}).$$

They act as one atom with scattering factor $m \times f$ rather than as m different atoms, each with scattering factor f. According to equation (2.1.4.11), this increases $F(\mathbf{h})$ by a factor $m^{1/2}$ on average. To make the F values of all reflections statistically comparable, $F(\mathbf{h})$ must be divided by $m^{1/2}$. For a detailed discussion, see *IT* B (2008), Chapter 2.1, by U. Shmueli and A. J. C. Wilson.

2.1.5. Reciprocal space and the Ewald sphere

A most convenient tool in X-ray crystallography is the reciprocal lattice. Unlike real or direct space, reciprocal space is imaginary. The reciprocal lattice is a superior instrument for constructing the X-ray diffraction pattern, and it will be introduced in the following way. Remember that vector $\mathbf{S}(hkl)$ is perpendicular to a reflecting plane and has a length $|\mathbf{S}(hkl)| = 2\sin\theta/\lambda = 1/d(hkl)$ (Section 2.1.4.5). This will now be applied to the boundary planes of the unit cell: the *bc* plane or (100), the *ac* plane or (010) and the *ab* plane or (001).

For the *bc* plane or (100): indices $h = 1$, $k = 0$ and $l = 0$; $\mathbf{S}(100)$ is normal to this plane and has a length $1/d(100)$. Vector $\mathbf{S}(100)$ will be called $\mathbf{a}^*$.

For the *ac* plane or (010): indices $h = 0$, $k = 1$ and $l = 0$; $\mathbf{S}(010)$ is normal to this plane and has a length $1/d(010)$. Vector $\mathbf{S}(010)$ will be called $\mathbf{b}^*$.

For the *ab* plane or (001): indices $h = 0$, $k = 0$ and $l = 1$; $\mathbf{S}(001)$ is normal to this plane and has a length $1/d(001)$. Vector $\mathbf{S}(001)$ will be called $\mathbf{c}^*$.

From the definition of $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ and the Laue conditions [equation (2.1.4.7)], the following properties of the vectors $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ can be derived:

$$\mathbf{a}^* \cdot \mathbf{a} = \mathbf{a} \cdot \mathbf{a}^* = \mathbf{a} \cdot \mathbf{S}(100) = h = 1.$$

Similarly

$$\mathbf{b}^* \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{S}(010) = k = 1,$$

and

$$\mathbf{c}^* \cdot \mathbf{c} = \mathbf{c} \cdot \mathbf{S}(001) = l = 1.$$

However, $\mathbf{a}^* \cdot \mathbf{b} = 0$ and $\mathbf{a}^* \cdot \mathbf{c} = 0$ because $\mathbf{a}^*$ is perpendicular to the (100) plane, which contains the *b* and *c* axes. Correspondingly, $\mathbf{b}^* \cdot \mathbf{a} = \mathbf{b}^* \cdot \mathbf{c} = 0$ and $\mathbf{c}^* \cdot \mathbf{a} = \mathbf{c}^* \cdot \mathbf{b} = 0$.

Proposition. The endpoints of the vectors $\mathbf{S}(hkl)$ form the points of a lattice constructed with the unit vectors $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$.

Proof. Vector $\mathbf{S}$ can be split into its coordinates along the three directions a^*, b^* and c^*:

$$\mathbf{S} = X \cdot \mathbf{a}^* + Y \cdot \mathbf{b}^* + Z \cdot \mathbf{c}^*. \tag{2.1.5.1}$$

Our proposition is true if X, Y and Z are whole numbers and indeed they are. Multiply equation (2.1.5.1) on the left and right side by $\mathbf{a}$.

$$\begin{array}{ccccccc} \mathbf{a} \cdot \mathbf{S} = & X \cdot \mathbf{a} \cdot \mathbf{a}^* & + & Y \cdot \mathbf{a} \cdot \mathbf{b}^* & + & Z \cdot \mathbf{a} \cdot \mathbf{c}^* \\ \vdots & \vdots & & \vdots & & \vdots \\ = h & = X \cdot 1 & & = 0 & & = 0. \end{array}$$

The conclusion is that $X = h$, $Y = k$ and $Z = l$, and, therefore,

$$\mathbf{S} = h \cdot \mathbf{a}^* + k \cdot \mathbf{b}^* + l \cdot \mathbf{c}^*.$$

The diffraction by a crystal [equation (2.1.4.6)] is only different from zero if the Laue conditions [equation (2.1.4.7)] are satisfied. All vectors $\mathbf{S}(hkl)$ are vectors in reciprocal space ending in reciprocal-lattice points and not in between. Each vector $\mathbf{S}(hkl)$ is normal to the set of planes (hkl) in real space and has a length $1/d(hkl)$ (Fig. 2.1.5.1).

The reciprocal-lattice concept is most useful in constructing the directions of diffraction. The procedure is as follows:

Step 1: Draw the vector $\mathbf{s}_o$ indicating the direction of the incident beam from a point M to the origin, O, of the reciprocal lattice. As in Section 2.1.4.2, the length of $\mathbf{s}_o$ and thus the distance MO is $1/\lambda$ (Fig. 2.1.5.2).

Step 2: Construct a sphere with radius $1/\lambda$ and centre M. The sphere is called the Ewald sphere. The scattering object is thought to be placed at M.

Step 3: Move a reciprocal-lattice point P to the surface of the sphere. Reflection occurs with $\mathbf{s} = \mathbf{MP}$ as the reflected beam, but only if the reciprocal-lattice point P is on the surface of the sphere, because only then does $\mathbf{S}(hkl) = \mathbf{s} - \mathbf{s}_o$ (Section 2.1.4.2).

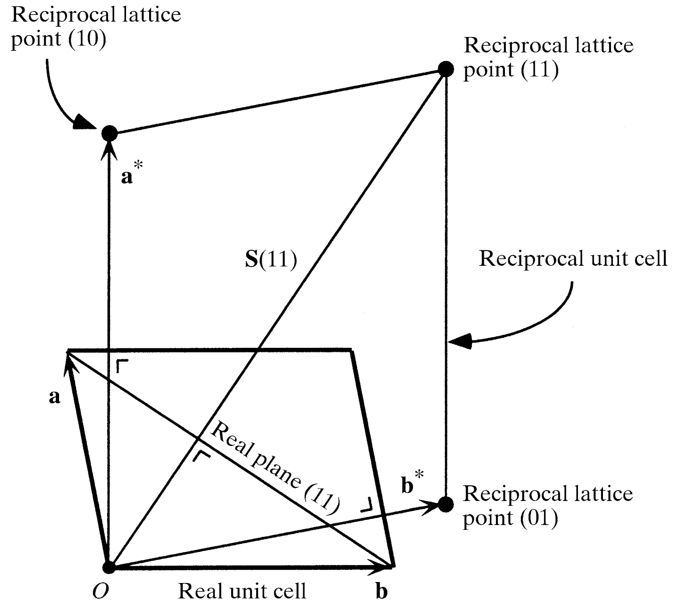

Reciprocal lattice point (10)
Reciprocal lattice point (11)
a*
S(11)
Reciprocal unit cell
a
Real plane (11)
b*
Reciprocal lattice point (01)
O
Real unit cell
b

Figure 2.1.5.1
A two-dimensional real unit cell is drawn together with its reciprocal unit cell. The reciprocal-lattice points are the endpoints of the vectors $\mathbf{S}(hk)$ [in three dimensions $\mathbf{S}(hkl)$]; for instance, vector $\mathbf{S}(11)$ starts at O and ends at reciprocal-lattice point (11). Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

Noncrystalline objects scatter differently. Their scattered waves are not restricted to reciprocal-lattice points passing through the Ewald sphere. They scatter in all directions.

2.1.6. Mosaicity and integrated reflection intensity

Crystals hardly ever have a perfect arrangement of their molecules, and crystals of macromolecules are certainly not perfect. Their crystal lattices show defects, which can sometimes be observed with an atomic force microscope or by interferometry. A schematic but useful way of looking at non-perfect crystals is through mosaicity; the crystal consists of a large number of tiny blocks. Each block is regarded as a perfect crystal, but the blocks are slightly misaligned with respect to each other. Scattering from different blocks is incoherent. Mosaicity causes a spread in the diffracted beams; when combined with the divergence of the beam from the X-ray source, this is called the effective mosaic spread. For the same crystal, effective mosaicity is smaller in a synchrotron beam with its lower divergence than in the laboratory. Protein crystals usually show a mosaic spread of 0.25–0.5°. Mosaic spread increases due to distortion of the lattice; this can happen as a result of flash freezing or radiation damage, for instance.

In Section 2.1.4.5, it was stated that the amplitude of the wave scattered by a crystal is proportional to the structure-factor amplitude $|F|$ and that its intensity is proportional to $|F|^2$. Of course, other factors also determine the intensity of the scattered beam, such as the wavelength, the intensity of the incident beam, the volume of the crystal *etc.* The intensity integrated over the entire region of the diffraction spot *hkl* is

$$I_{\rm int}(hkl) = \frac{\lambda^3}{\omega V^2}\left(\frac{e^2}{mc^2}\right)^2 V_{\rm cr} I_o LPT|F(hkl)|^2. \tag{2.1.6.1}$$

In equation (2.1.6.1), we recognize $I_o(e^2/mc^2)^2$ as part of the Thomson scattering for one electron, $I_{\rm el} = I_o(e^2/mc^2)^2 \sin^2\varphi$ [equations (2.1.4.1*a*) and (2.1.4.1*b*)] per unit solid angle. $V_{\rm cr}$ is the

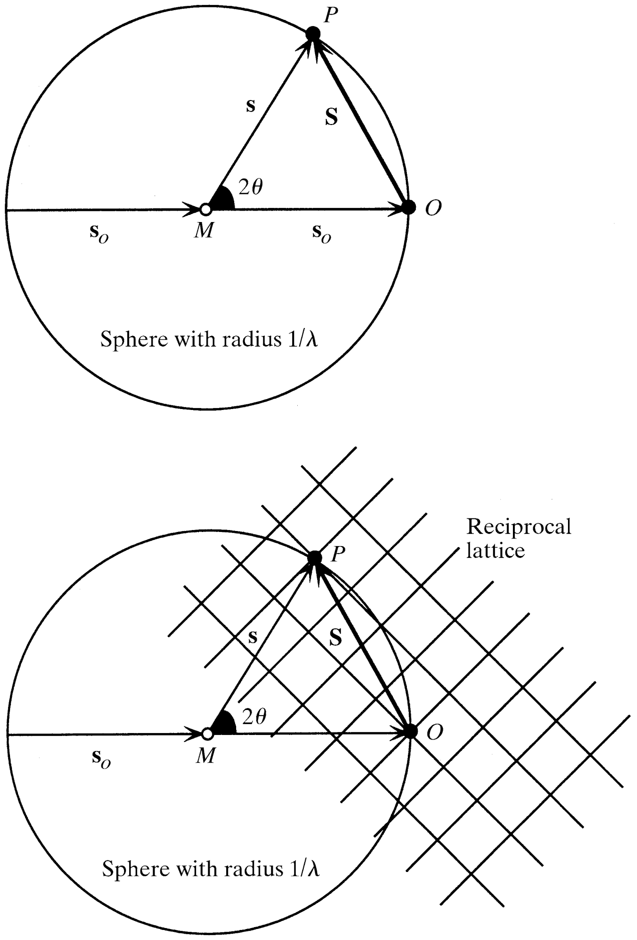


Figure 2.1.5.2
The circle is, in fact, a sphere with radius $1/\lambda$. $\mathbf{s}_O$ indicates the direction of the incident beam and has a length $1/\lambda$. The diffracted beam is indicated by vector **s**, which also has a length $1/\lambda$. Only reciprocal-lattice points on the surface of the sphere are in a reflecting position. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

volume of the crystal and V is the volume of the unit cell. It is clear that the scattered intensity is proportional to the volume of the crystal. The term $1/V^2$ can be explained as follows. In a mosaic block, all unit cells scatter in phase. For a given volume of the individual blocks, the number of unit cells in a mosaic block, as well as the scattering amplitude, is proportional to $1/V$. The scattered intensity is then proportional to $1/V^2$. Because of the finite reflection width, scattering occurs not only for the reciprocal-lattice point when it is on the Ewald sphere, but also for a small volume around it. Since the sphere has radius $1/\lambda$, the solid angle for scattering, and thus the intensity, is proportional to $1/(1/\lambda)^2 = \lambda^2$.

However, in equation (2.1.6.1), the scattered intensity is proportional to λ^3. The extra λ dependence is related to the time t it takes for the reciprocal-lattice 'point' to pass through the surface of the Ewald sphere. With an angular speed of rotation ω, a reciprocal-lattice point at a distance $1/d$ from the origin of the reciprocal lattice moves with a linear speed $v = (1/d)\omega$ if the rotation axis is normal to the plane containing the incident and reflected beam. For the actual passage through the surface of the Ewald sphere, the component perpendicular to the surface is needed: $v_\perp = (1/d)\omega\cos\theta = \omega\sin 2\theta/\lambda$. Therefore, the time t required to pass through the surface is proportional to $(1/\omega)(\lambda/\sin 2\theta)$. This introduces the extra λ term in equation (2.1.6.1) as well as the ω dependence and a $1/\sin 2\theta$ term. The latter represents the Lorentz factor L. It is a geometric correction factor for the *hkl* reflections; here it is $1/\sin 2\theta$, but it is different for other data-collection geometries.

The factor P in equation (2.1.6.1) is the polarization factor. For the polarized incident beam used in deriving equation (2.1.4.1*a*), $P = \sin^2\varphi$, where φ is the angle between the polarization direction of the beam and the scattering direction. It is easy to verify that $\phi = 90^\circ - 2\theta$, where θ is the reflecting angle (Fig. 2.1.4.9). P depends on the degree of polarization of the incident beam. For a completely unpolarized beam, $P = (1 + \cos^2 2\theta)/2$.

In equation (2.1.6.1), T is the transmission factor: $T = 1 - A$, where A is the absorption factor. When X-rays travel through matter, they suffer absorption. The overall absorption follows Beer's law:

$$I = I_o \exp(-\mu d),$$

where I_o is the intensity of the incident beam, d is the path length in the material and μ is the total linear absorption coefficient. μ can be obtained as the sum of the atomic mass absorption coefficients of the elements $(\mu_m)_i$:

$$\mu = \rho \sum_i g_i(\mu_m)_i,$$

where ρ is the density of the absorbing material and g_i is the mass fraction of element i.

Atomic mass absorption coefficients $(\mu_m)_i$ for the elements are listed in Tables 4.2.4.3 (and 4.2.4.1) of *IT* C (2004) as a function of a large number of wavelengths. The absorption is wavelength-dependent and is generally much stronger for longer wavelengths. This is the result of several processes. For the X-ray wavelengths applied in crystallography, the processes are scattering and photoelectric absorption. Moreover, at the reflection position, the intensity may be reduced by extinction.

Scattering is the result of a collision between the X-ray photons and the electrons. One can distinguish two kinds of scattering: Compton scattering and Rayleigh scattering. In Compton scattering, the photons lose part of their energy in the collision process (inelastic scattering), resulting in scattered photons with a lower energy and a longer wavelength. Compton scattering contributes to the background in an X-ray diffraction experiment. In Rayleigh scattering, the photons are elastically scattered, do not lose energy, and leave the material with their wavelength unchanged. In a crystal, they interfere with each other and give rise to the Bragg reflections. Between the Bragg reflections, there is no loss of energy due to elastic scattering and the incident beam is hardly reduced. In the Bragg positions, if the reduction in intensity of the incident beam due to elastic scattering can still be neglected, the crystal is considered an ideal mosaic. For non-ideal mosaic crystals, the beam intensity is reduced by *extinction*:

(1) The blocks are too large, and multiple reflection occurs within a block. At each reflection process, the phase angle shifts $\pi/2$ (Section 2.1.4.3.2). After two reflections, the beam travels in the same direction as the incident beam but with a phase difference of π, and this reduces the intensity.
(2) The angular spread of the blocks is too small. The incident beam is partly reflected by blocks close to the surface and the resulting beam is the incident beam for the lower-lying blocks that are also in reflecting position.

Extinction is not a serious problem in protein X-ray crystallography.

Absorption curves as a function of the X-ray wavelength show anomalies at absorption edges. At such an edge, electrons are ejected from the atom or are elevated to a higher-energy bound state, the photons disappear completely and the X-ray beam is strongly absorbed. This is called *photoelectric absorption*. At an absorption edge, the frequency of the X-ray beam ν is equal to the frequency ν_K, ν_L or ν_M corresponding to the energy of the K, L or M state. According to equation (2.1.4.4), anomalous scattering is maximal at an absorption edge.

2.1.7. Calculation of electron density

In equation (2.1.4.6), the wave $W_{\rm cr}(\mathbf{S})$ scattered by the crystal is given as the sum of the atomic contributions, as in equation (2.1.4.5) for the scattering by a unit cell. In the derivation of equation (2.1.4.5), it is assumed that the atoms are spherically symmetric (Section 2.1.4.3) and that density changes due to chemical bonding are neglected. A more exact expression for the wave scattered by a crystal, in the absence of anomalous scattering, is

$$W_{\rm cr}(\mathbf{S}) = \int_{\rm crystal} \rho(\mathbf{r}) \exp(2\pi i \mathbf{r}\cdot\mathbf{S})\, {\rm d}v_{\rm real}. \tag{2.1.7.1}$$

The integration is over all electrons in the crystal. $\rho(\mathbf{r})$ is the electron-density distribution in each unit cell. The operation on the electron-density distribution in equation (2.1.7.1) is called Fourier transformation, and $W_{\rm cr}(\mathbf{S})$ is the Fourier transform of $\rho(\mathbf{r})$. It can be shown that $\rho(\mathbf{r})$ is obtained by an inverse Fourier transformation:

$$\rho(\mathbf{r}) = \int_{\mathbf{S}} W_{\rm cr}(\mathbf{S}) \exp(-2\pi i \mathbf{r}\cdot\mathbf{S})\, {\rm d}v_{\rm reciprocal}. \tag{2.1.7.2}$$

In contrast to $\rho(\mathbf{r})$, $W_{\rm cr}(\mathbf{S})$ is not a continuous function but, because of the Laue conditions, it is only different from zero at the reciprocal-lattice points $\mathbf{h}$ $(= hkl)$. In equation (2.1.4.6), $W_{\rm cr}(\mathbf{S})$ is the product of the structure factor and three delta functions. The structure factor at the reciprocal-lattice points is $F(\mathbf{h})$, and the product of the three delta functions is $1/V$, the volume of one reciprocal unit cell. Therefore, $W_{\rm cr}(\mathbf{S})$ in equation (2.1.7.2) can be replaced by $F(\mathbf{h})/V$, and equation (2.1.7.2) itself by

$$\rho(\mathbf{r}) = (1/V) \sum_{\mathbf{h}} F(\mathbf{h}) \exp(-2\pi i \mathbf{r}\cdot\mathbf{h}). \tag{2.1.7.3}$$

If x, y and z are fractional coordinates in the unit cell, $\mathbf{r}\cdot\mathbf{S} = (\mathbf{a}\cdot x + \mathbf{b}\cdot y + \mathbf{c}\cdot z)\cdot\mathbf{S} = \mathbf{a}\cdot\mathbf{S}\cdot x + \mathbf{b}\cdot\mathbf{S}\cdot y + \mathbf{c}\cdot\mathbf{S}\cdot z = hx + ky + lz$, and an alternative expression for the electron density is

$$\rho(xyz) = (1/V) \sum_h \sum_k \sum_l F(hkl) \exp[-2\pi i(hx + ky + lz)]. \tag{2.1.7.4}$$

Instead of expressing $F(\mathbf{S})$ as a summation over the atoms [equation (2.1.4.5)], it can be expressed as an integration over the electron density in the unit cell:

$$F(hkl) = V \int_{x=0}^{1} \int_{y=0}^{1} \int_{z=0}^{1} \rho(xyz) \exp[2\pi i(hx + ky + lz)]\, {\rm d}x\, {\rm d}y\, {\rm d}z. \tag{2.1.7.5}$$

Because $F(hkl)$ is a vector in the Argand diagram with an amplitude $|F(hkl)|$ and a phase angle $\alpha(hkl)$,

$$F(hkl) = |F(hkl)| \exp[i\alpha(hkl)]$$

and

$$\rho(xyz) = (1/V) \sum_h \sum_k \sum_l |F(hkl)| \exp[-2\pi i(hx + ky + lz) + i\alpha(hkl)]. \tag{2.1.7.6}$$

By applying equation (2.1.7.6), the electron-density distribution in the unit cell can be calculated, provided values of $|F(hkl)|$ and $\alpha(hkl)$ are known. From equation (2.1.6.1), it is clear that $|F(hkl)|$ can be derived, on a relative scale, from $I_{\rm int}(hkl)$ after a correction for the background and absorption, and after application of the Lorentz and polarization factor:

$$|F(hkl)| = \left[\frac{I_{\rm int}(hkl)}{LPT}\right]^{1/2}. \tag{2.1.7.7}$$

Contrary to the situation with crystals of small compounds, it is not easy to find the phase angles $\alpha(hkl)$ for crystals of macromolecules by direct methods, although these methods are in a state of development (see Part 16). Indirect methods to determine the protein phase angles are:

(1) isomorphous replacement (see Part 12);
(2) molecular replacement (see Part 13);
(3) multiple-wavelength anomalous dispersion (MAD) (see Part 14).

From equation (2.1.7.5), it is clear that the reflections hkl and $\overline{hkl}$ have the same value for their structure-factor amplitudes, $|F(hkl)| = |F(\overline{hkl})|$, and for their intensities, $I(hkl) = I(\overline{hkl})$, but have opposite values for their phase angles, $\alpha(hkl) = -\alpha(\overline{hkl})$, assuming that anomalous dispersion can be neglected. Consequently, equation (2.1.7.6) reduces to

$$\rho(xyz) = (1/V) \sum_h \sum_k \sum_l |F(hkl)| \cos[2\pi(hx + ky + lz) - \alpha(hkl)] \tag{2.1.7.8}$$

or

$$\rho(xyz) = F(000)/V + (2/V) {\sum_h}' {\sum_k}' {\sum_l}' |F(hkl)| \times \cos[2\pi(hx + ky + lz) - \alpha(hkl)]. \tag{2.1.7.9}$$

$\sum'$ denotes that $F(000)$ is excluded from the summation and that only the reflections hkl, and not $\overline{hkl}$, are considered.

The two reflections, hkl and $\overline{hkl}$, are called Friedel or Bijvoet pairs.

If anomalous dispersion cannot be neglected, the two members of a Friedel pair have different values for their structure-factor amplitudes, and their phase angles no longer have opposite values. This is caused by the f'' contribution to the anomalous scattering (Fig. 2.1.7.1). Macromolecular crystals show anomalous dispersion if the structure contains, besides the light atoms, one or more heavier atoms. These can be present in the native structure or are introduced in the isomorphous replacement technique or in MAD analysis.

2.1.8. Symmetry in the diffraction pattern

In the previous section, it was noted that $I(hkl) = I(\overline{hkl})$ if anomalous scattering can be neglected. In this case, the effect is that the diffraction pattern has a centre of symmetry. This is also true for the reciprocal lattice if the reciprocal-lattice points (hkl) are weighted with their $I(hkl)$ values. If the crystal structure has symmetry elements, they are also found in the diffraction pattern and in the weighted reciprocal lattice. Macromolecular crystals of biological origin are enantiomorphic and the symmetry operators in the crystal are restricted to rotation axes and screw axes. It is evident that a rotation of the real lattice will cause the same

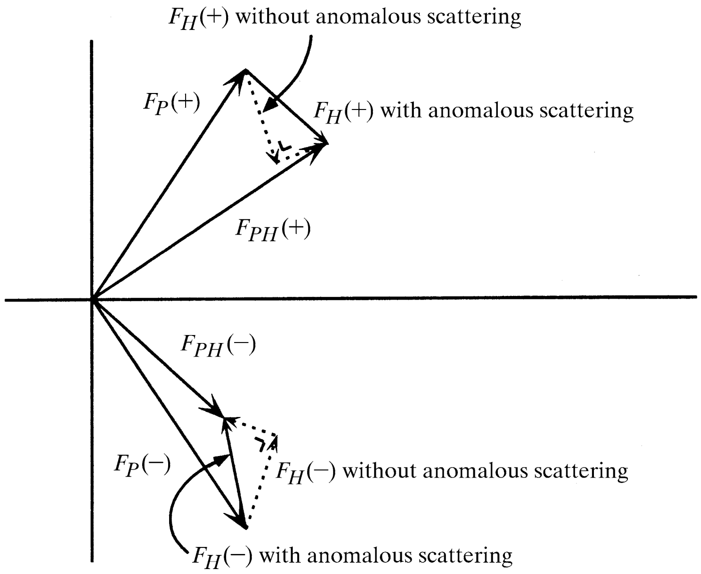

$F_H(+)$ without anomalous scattering
$F_P(+)$
$F_H(+)$ with anomalous scattering
$F_{PH}(+)$
$F_{PH}(-)$
$F_P(-)$
$F_H(-)$ without anomalous scattering
$F_H(-)$ with anomalous scattering

Figure 2.1.7.1
An Argand diagram for the structure factors of the two members of a Friedel pair. (+) represents *hkl* and (−) represents $\overline{hkl}$. F_P is the contribution to the structure factor by the non-anomalously scattering protein atoms and F_H is that for the anomalously scattering atoms. F_H consists of a real part with an imaginary part perpendicular to it. The real parts are mirror images with respect to the horizontal axis. The imaginary parts are rotated counterclockwise with respect to the real parts (Section 2.1.4.4). The result is that the total structure factors, $F_{PH}(+)$ and $F_{PH}(-)$, have different amplitudes and phase angles. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

rotation of the reciprocal lattice. If this rotation is the result of a symmetry operation around an axis, the crystal structure looks exactly the same as before the rotation, and the same must be true for the weighted reciprocal lattice. However, screw axes in the crystal lattice reduce to normal (non-screw) rotation axes in the weighted reciprocal lattice, as has been shown by Waser (1955). We follow his arguments, but must first introduce matrix notation for convenience.

If **r** is a position vector and **h** a vector in reciprocal space, the scalar product

$$\mathbf{h}\cdot\mathbf{r} = (h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*)\cdot(\mathbf{a}x + \mathbf{b}y + \mathbf{c}z) = hx + ky + lz,$$

or in matrix notation,

$$(hkl)\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{h}^T\mathbf{r},$$

where $(hkl) = \mathbf{h}^T$ is a row vector and

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{r}$$

is a column vector. $\mathbf{h}^T$ is the transpose of column vector **h** (rows and columns are interchanged). In this notation, the structure factor is given by

$$F(\mathbf{h}) = \int_{\text{cell}} \rho(\mathbf{r})\exp(2\pi i\mathbf{h}^T\cdot\mathbf{r})\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.1)$$

The symmetry operation of a screw axis is a combination of a rotation and a translation. The rotation can be represented by the matrix **R** and the translation by the vector **t**. Because of the screw-axis symmetry, $\rho(\mathbf{R}\cdot\mathbf{r}+\mathbf{t}) = \rho(\mathbf{r})$.

$F(\mathbf{h})$ can also be expressed as

$$\begin{aligned} F(\mathbf{h}) &= \int_{\text{cell}} \rho(\mathbf{R}\cdot\mathbf{r}+\mathbf{t})\exp[2\pi i\mathbf{h}^T\cdot(\mathbf{R}\cdot\mathbf{r}+\mathbf{t})]\,\mathrm{d}v_{\text{real}} \\ &= \exp(2\pi i\mathbf{h}^T\cdot\mathbf{t})\int_{\text{cell}} \rho(\mathbf{r})\exp(2\pi i\mathbf{h}^T\cdot\mathbf{R}\cdot\mathbf{r})\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.2) \end{aligned}$$

Because $\mathbf{h}^T\cdot\mathbf{R} = (\mathbf{R}^T\cdot\mathbf{h})^T$, where $\mathbf{R}^T$ is the transpose of the matrix **R**, equation (2.1.8.2) can be written as

$$F(\mathbf{h}) = \exp(2\pi i\mathbf{h}^T\cdot\mathbf{t})\int_{\text{cell}} \rho(\mathbf{r})\exp[2\pi i(\mathbf{R}^T\cdot\mathbf{h})^T\cdot\mathbf{r}]\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.3)$$

By definition, the integral in equation (2.1.8.3) is $F\,(\mathbf{R}^T\cdot\mathbf{h})$, and, therefore

$$F(\mathbf{h}) = \exp(2\pi i\mathbf{h}^T\cdot\mathbf{t})F(\mathbf{R}^T\cdot\mathbf{h}).$$

Conclusion: The phase angles of the two structure factors are different for $\mathbf{t}\neq 0$:

$$\alpha(\mathbf{h}) = \alpha(\mathbf{R}^T\cdot\mathbf{h}) + 2\pi\mathbf{h}^T\cdot\mathbf{t}, \qquad (2.1.8.4)$$

but the structure-factor amplitudes and, therefore, the intensities are always equal:

$$I(\mathbf{h}) = I(\mathbf{R}^T\cdot\mathbf{h}) \quad\text{or}\quad I[(\mathbf{R}^T)^{-1}\cdot\mathbf{h}] = I(\mathbf{h}). \qquad (2.1.8.5)$$

The matrices $(\mathbf{R}^T)^{-1}$ in reciprocal space and **R** in direct space denote rotation over the same angle. Therefore, both an *n*-fold screw axis and an *n*-fold rotation axis in the crystal correspond to an *n*-fold axis in the weighted reciprocal lattice.

However, screw axes distinguish themselves from non-screw axes by extinction of some reflections along the line in reciprocal space corresponding to the screw-axis direction. This will be shown for a twofold screw axis along the monoclinic *b* axis.

The electron density at **r**, $\rho(\mathbf{r})$, is then equal to the electron density at $\mathbf{R}\cdot\mathbf{r}+\mathbf{t}$, where $\mathbf{R}\cdot\mathbf{r}$ is a rotation that leaves the value of the *y* coordinate unchanged. **t** is equal to **b**/2.

$$F(\mathbf{h}) = \int_{\text{half the cell}} \rho(\mathbf{r})\{\exp(2\pi i\mathbf{h}^T\cdot\mathbf{r}) + \exp[2\pi i\mathbf{h}^T(\mathbf{R}\cdot\mathbf{r}+\mathbf{t})]\}\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.6)$$

For the (0*k*0) reflections, (**h** along $\mathbf{b}^*$) is $\mathbf{h} = k\mathbf{b}^*$, giving

$$\mathbf{h}^T\cdot\mathbf{r} = \mathbf{h}^T\cdot\mathbf{R}\cdot\mathbf{r} = 0 + ky + 0 \text{ and } \mathbf{h}^T\cdot\mathbf{t} = k/2.$$

This simplifies equation (2.1.8.6) to

$$F(0k0) = [1+\exp(\pi ik)]\int_{\text{half the cell}} \rho(\mathbf{r})\exp(2\pi iky)\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.7)$$

If *k* is odd, $F(0k0) = 0$, because $1+\exp(\pi ik) = 0$.

This type of systematic absence, due to screw components in the symmetry elements, occurs along lines in reciprocal space. Other types of absence apply to all *hkl* reflections. They result from the centring of the unit cell (Fig. 2.1.1.4). Suppose the unit cell is centred in the *ab* plane (*C* centring). Consequently, the electron density at **r** is equal to the electron density at $\mathbf{r}+\mathbf{t}$, with $\mathbf{t} = \mathbf{a}/2 + \mathbf{b}/2$ and $\mathbf{h}^T\cdot\mathbf{t} = h/2 + k/2$. The structure factor can then be written as

$$F(\mathbf{h}) = \{1+\exp[\pi i(h+k)]\}\int_{\text{half the cell}} \rho(\mathbf{r})\exp\big(2\pi i\mathbf{h}^T\cdot\mathbf{r}\big)\,\mathrm{d}v_{\text{real}}. \qquad (2.1.8.8)$$

The conclusion is that when $(h+k)$ is odd, the structure factors are zero and no diffracted intensity is observed for those reflections.

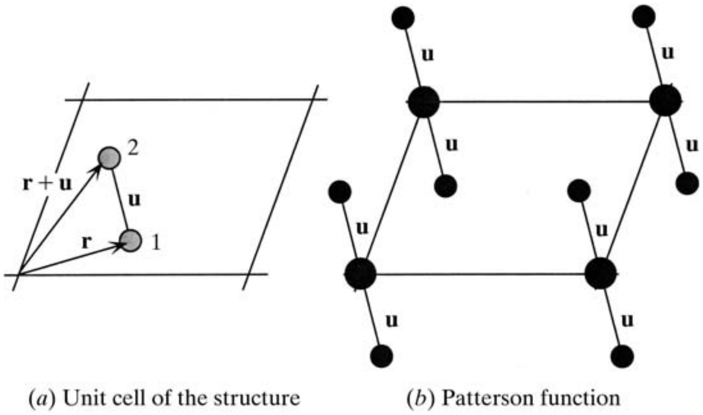


Figure 2.1.9.1
(*a*) A two-dimensional unit cell with two atoms. (*b*) The corresponding Patterson function. Reproduced with permission from Drenth (1999). Copyright (1999) Springer-Verlag.

2.1.9. The Patterson function

In 1934, A. L. Patterson presented a method for locating the atomic positions in not too complicated molecules without knowledge of the phase angles (Patterson, 1934). The method involves the calculation of the Patterson function, $P(uvw) = P(\mathbf{u})$:

$$P(\mathbf{u}) = (1/V)\sum_{\mathbf{h}} |F(\mathbf{h})|^2 \cos(2\pi\mathbf{h}\cdot\mathbf{u}), \qquad (2.1.9.1)$$

or, written as an exponential function,

$$P(\mathbf{u}) = (1/V)\sum_{\mathbf{h}} |F(\mathbf{h})|^2 \exp(2\pi i\mathbf{h}\cdot\mathbf{u}). \qquad (2.1.9.2)$$

Equations (2.1.9.1) and (2.1.9.2) give the same result, because in the definition of $P(\mathbf{u})$ anomalous dispersion is neglected, resulting in $|F(\mathbf{h})|^2 = |F(-\mathbf{h})|^2$. Comparison with equations (2.1.7.3) and (2.1.7.6) shows that the Patterson function $P(\mathbf{u})$ is a Fourier summation with coefficients $|F(\mathbf{h})|^2$ instead of $F(\mathbf{h}) = |F(\mathbf{h})|\exp[i\alpha(\mathbf{h})]$. The periodicity, and thus the unit cell, are the same for the electron density and the Patterson function. For the Patterson function, many authors prefer to use **u** rather than **r** as the position vector.

The fundamental advantage of Patterson's discovery is that, in contrast to the calculation of $\rho(\mathbf{r})$, no phase information is needed for calculating $P(\mathbf{u})$.

The Patterson map can be obtained directly after the intensities of the reflections have been measured and corrected. However, what kind of information does it provide? This can be understood from an alternative expression for the Patterson function:

$$P(\mathbf{u}) = \int_{\mathbf{r}} \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\,\mathrm{d}v_{\text{real}}. \qquad (2.1.9.3)$$

Equation (2.1.9.3) leads to the same result as equation (2.1.9.1), as can be proved easily by substituting expression (2.1.7.3) for ρ in the right-hand side of equation (2.1.9.3).

On the right-hand side of the equation, the electron density $\rho(\mathbf{r})$ at position **r** in the unit cell is multiplied by the electron density $\rho(\mathbf{r}+\mathbf{u})$ at position $\mathbf{r}+\mathbf{u}$; the integration is over all vectors **r** in the unit cell. The result of the integration is that the Patterson map will show peaks at the end of vectors **u** between atoms in the unit cell of the structure; all these Patterson vectors start at the origin of the Patterson cell. This can best be understood with a simple example. In Fig. 2.1.9.1, a two-dimensional unit cell is drawn containing only two atoms (1 and 2). To calculate the Patterson map, a vector **u** must be moved through this cell, and, according to equation (2.1.9.3), for every position and orientation of **u**, the electron densities at the beginning and at the end of **u** must be multiplied. It is clear that this product will generally be zero unless the length and the orientation of **u** are such that it begins in atom 1 and ends in atom 2, or the other way around. If so, there is a peak in the Patterson map at the end of vector **u** and at the end of vector $-\mathbf{u}$, implying that the Patterson map is always centrosymmetric. The origin itself, where vector $\mathbf{u} = 0$, always has a high peak because

$$P(\mathbf{u}=0) = \int_{\mathbf{r}} \rho(\mathbf{r})\rho(\mathbf{r})\,\mathrm{d}v_{\text{real}} = \sum_{\mathbf{h}} |F(\mathbf{h})|^2.$$

The origin peak is equal to the sum of the squared local electron densities. The height of each non-origin peak is proportional to the product of $\rho(\mathbf{r})$ and $\rho(\mathbf{r}+\mathbf{u})$. This is an important feature in the isomorphous replacement method for protein-structure determination, in which the heavy-atom positions are derived from a difference Patterson calculated with coefficients $(|F_{PH}| - |F_P|)^2$, where $|F_{PH}|$ is the structure-factor amplitude of the heavy-atom derivative and $|F_P|$ is that of the native protein (see Part 12). The vectors between the heavy atoms are the most prominent features in such a map.

The number of peaks in a Patterson map increases much faster than the number of atoms. For n atoms in the real unit cell, there are n^2 Patterson peaks, n of them superimposed at the origin, and $n \times (n-1)$ elsewhere in the Patterson cell. Because the atomic electron densities cover a certain region and the width of a Patterson peak at **u** is roughly the sum of the widths of the atoms connected by **u**, overlap of peaks is a real problem in the interpretation of a Patterson map. It can almost only be done for unit cells with a restricted number of atoms unless some extra information is available. For crystals of macromolecules, it is certainly impossible to derive the structure from an interpretation of the Patterson map.

The situation can be improved through sharpening the Patterson peaks by simulating the atoms as point scatterers. This can be achieved by replacing the $|F(\mathbf{h})|^2$ values with modified intensities which, on average, do not decrease with $\sin\theta/\lambda$. For instance, suitable intensities for this purpose are the squared normalized structure-factor amplitudes $|E(\mathbf{h})|^2$ (Section 2.1.4.6), the average of which is 1 at all $\sin\theta/\lambda$. A disadvantage of sharpening to point peaks is the occurrence of diffraction ripples around the sharp peaks, induced by truncation of the Fourier series in equation (2.1.9.1). Therefore, modified intensities corresponding to less sharpened peaks are sometimes used [*IT* B (2008), Chapter 2.3, pp. 245–246]. Diffraction ripples that seriously disturb the Patterson map are generated by the high origin peak, and, particularly for sharpened maps, it is advisable to remove this peak. This implies that $P(\mathbf{u}=0) = 0$ [equation (2.1.9.1)]. It is easy to verify that this requires coefficients $[|F(\mathbf{h})|^2 - \langle|F(\mathbf{h})|^2\rangle]$ for the $|F(\mathbf{h})|^2$ map and $[|E(\mathbf{h})|^2 - 1]$ for the $|E(\mathbf{h})|^2$ map. Note that the term for $\mathbf{h} = 0$ is omitted and that the average of $|F(\mathbf{h})|^2$ must be taken for the appropriate $\sin\theta/\lambda$ region.

The symmetry in a Patterson map is related to the symmetry in the electron-density map, but it is not necessarily the same. For instance, screw axes in the real cell become non-screw axes in the Patterson cell, because all interatomic vectors start at the origin. It is possible, however, to distinguish between screw axes and non-screw axes by the concentration of peaks in the Patterson map. For instance, the consequence of a twofold symmetry axis along **b** is the presence of a large number of peaks in the $(u0w)$ plane of the Patterson map. For a screw axis with translation $\frac{1}{2}$ along **b**, the peaks lie in the $(u\frac{1}{2}w)$ plane. Such planes are called Harker planes (Harker, 1936). Peaks in Harker planes usually form the start of the interpretation of a Patterson map. Harker lines result from mirror planes, which do not occur in macromolecular crystal structures of biological origin.

Despite the improvements that can be made to the Patterson function, for structures containing atoms of nearly equal weight its complete interpretation can only be achieved for a restricted number of atoms per cell unless some extra information is available. Nowadays, most structure determinations of small compounds are based on direct methods for phase determination. However, these may fail for structures showing strong regularity. In these cases, Patterson interpretation is used as an alternative tool, sometimes in combination with direct methods. It is interesting to see that the value of the Patterson function has shifted from the small-compound field to macromolecular crystallography, where it plays an extremely useful role:

(1) in the isomorphous replacement method, the positions of the very limited number of heavy atoms attached to the macromolecule can be derived from a difference Patterson map, as mentioned earlier in this section;

(2) anomalous scatterers can be located by calculating a Patterson map with coefficients $[|F_{PH}(\mathbf{h})| - |F_{PH}(-\mathbf{h})|]^2$, in which $|F_{PH}(\mathbf{h})|$ is the structure-factor amplitude of the protein containing the anomalous scatterer;

(3) molecular replacement is based on the property that the Patterson map is a map of vectors between atoms in the real structure, combined with the fact that such a vector map is (apart from a rotation) similar for two homologous structures: the unknown and a known model structure.

I am greatly indebted to Aafje Looyenga-Vos for critically reading the manuscript and for many useful suggestions.

References

Burzlaff, H. & Zimmermann, H. (2005). *Bases, lattices, Bravais lattices and other classifications*. In *International Tables for Crystallography*, Vol. A. *Space-Group Symmetry*, edited by Th. Hahn, ch. 9.1. Heidelberg: Springer.

Drenth, J. (1999). *Principles of Protein X-ray Crystallography*. New York: Springer-Verlag.

Haas, C. & Drenth, J. (1995). *The interaction energy between two protein molecules related to physical properties of their solution and their crystals and implications for crystal growth*. *J. Cryst. Growth*, **154**, 126–135.

Harker, D. (1936). *The application of the three-dimensional Patterson method and the crystal structures of proustite, Ag_3AsS_3, and pyrargyrite, Ag_3SbS_3*. *J. Chem. Phys*. **4**, 381–390.

Heitler, W. G. (1966). *The Quantum Theory of Radiation*, 3rd ed. Oxford University Press.

Hönl, H. (1933). *Atomfaktor für Röntgenstrahlen als Problem der Dispersionstheorie (K-Schale)*. *Ann. Phys*. **18**, 625–655.

International Tables for Crystallography (2008). Vol. B. *Reciprocal Space*, edited by U. Shmueli, 3rd ed. Heidelberg: Springer.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

International Tables for Crystallography (2005). Vol. A. *Space-Group Symmetry*, edited by Th. Hahn. Heidelberg: Springer.

James, R. W. (1965). *The Optical Principles of the Diffraction of X-rays*, p. 135. London: G. Bell and Sons Ltd.

Kauzmann, W. (1957). *Quantum Chemistry*. New York: Academic Press.

Klein, O. & Nishina, Y. (1929). *Über die Streuung von Strahlung durch freie Elektronen nach der neuen relativistischen Quantendynamik von Dirac*. *Z. Phys*. **52**, 853–868.

Patterson, A. L. (1934). *A Fourier series method for the determination of the components of interatomic distances in crystals*. *Phys. Rev*. **46**, 372–376.

Waser, J. (1955). *Symmetry relations between structure factors*. *Acta Cryst*. **8**, 595.

Wilson, A. J. C. (1942). *Determination of absolute from relative X-ray intensity data*. *Nature (London)*, **150**, 151–152.

references

International Tables for Crystallography (2012). Vol. F, Chapter 2.2, pp. 64–74.

Chapter 2.2. Quality indicators in macromolecular crystallography: definitions and applications

H. M. Einspahr and M. S. Weiss

2.2.1. Introduction

The genesis of this chapter was a perceived need for a single location in the volume in which consensus definitions could be found for the many statistical indicators of quality or figures of merit that have been developed to monitor a macromolecular crystallography (MX) experiment and its final outcome, a model. The evolving experiment has generated a rich collection of *R* values, signal-to-noise indicators, correlation coefficients and other figures of merit. As improvements in data collection, processing and other aspects of the experiment continue, we can expect a continued evolution of new or improved indicators of quality with which to monitor the impact of those improvements.

This chapter, then, attempts to provide a comprehensive list of the indicators of quality currently in use and, for each indicator in the list, a precise definition that conforms to consensus interpretations of the literature and current practice. The authors acknowledge that useful indicators may have been missed in the sweep that produced this list, but hasten to point out that the generation of new indicators is an ongoing process with the newest ones often in obscurity for a period before their utility is recognized and adopted by experimenters. There is also a subset of the indicators in the list whose members stand out as universally accepted and crucial indicators that either every experimental description must include or that every experimenter should be familiar with. A summary of these is given at the end of this article in Table 2.2.11.1. The authors also acknowledge that there may be competing definitions for some indicators. Where they occur, these competing definitions will be pointed out and a discussion will be provided which, at a minimum, will attempt to clarify the differences, but will also, where possible, attempt to arbitrate or discriminate among opposing views. It should be noted that the scope of this chapter is primarily concerned with the crystallographic experiment. While quality indicators useful for model refinement are given in Sections 2.2.8 to 2.2.10, the validation of the refined model is covered more extensively in Part 21 of this volume.

Before proceeding with the quality indicators themselves, it is necessary to discuss a small collection of parameters whose values, typically left to the experimenter to fix, impact on virtually all the indicators of quality we present here. The reason for this wide impact is that they are used to determine which reflections are included in the final data set in question. In their simplest form, the effect is a limit, applied as a cutoff in intensity or resolution, beyond which reflections are excluded from consideration.

When a cutoff is applied based on intensities, the limiting value may be simply a fixed number, for example zero, or it may vary from reflection to reflection, for example some multiple of the standard uncertainty of the reflection intensity. Reflections with intensities below limiting values are excluded. Reflections excluded in this manner are often referred to as unobserved.

When the objective is to estimate electron-density distributions, the exclusion of intensity terms below very low limiting values is unlikely to have any significant negative impact and may offer a positive savings in computation time. This may not be the case, however, when the objective is refinement. Inclusion of reflections of low intensity, even zero, may have important positive results on the quality of the final model. True, some of these intensities may be poorly known or, if less than zero, physically unrealistic, but given the excellent quality of modern data-collection instrumentation and techniques, few intensities, if any, are without some reasonable estimate of their standard uncertainty, so that weighting procedures can be applied to modulate the impact of individual terms in the final result and judgments can be made about exclusion of intensity values that are truly improbable. With the application of proper weighting procedures, there seems to be little justification for exclusion of reflections from consideration based on intensity (except for the truly improbable mentioned above) when computing indicators of quality. As instrumentation and techniques continue to improve, this is a topic that merits continued attention and debate within the MX community in search of consensus best practices for handling weak reflections and for including them in estimators of quality.

While an acceptable estimate of the nominal resolution of a diffraction data set is widely considered to be a high-value indicator of data quality, assignment of a value to this limit is typically left to the experimenter and is thus prone to subjectivity. Some guidelines for estimation have emerged. One would set the nominal resolution based on the percentage of weak reflections above the limit so that reflections with intensities above that limit would be included. An example might be the resolution at which 70% of unique reflections have intensities above zero or above some multiple of their standard uncertainties. Another way to estimate the nominal resolution which has gained wide acceptance is based on the overall signal-to-noise value where, in its most popular expression, the limit value is the resolution at which the mean signal-to-noise ratio in the outer resolution shell falls to 2. Each of these estimation methods is susceptible to convolution with limits based on intensity leading to, for example, a limit set as the resolution at which the mean signal-to-noise ratio 'of observed reflections' falls to 2. The guidelines for estimating resolution limits – the methods to be applied, the constraining values such as 70% or 2, and the proper integration with limits applied based on intensity – also merit further attention and discussion by the MX community, the goal being definition of consensus best practices. We take this opportunity to suggest that the widely accepted estimate, the resolution at which the mean signal-to-noise in the outer resolution shell falls to 2, calculated without imposition of a cutoff in intensity, has much to recommend it. True, some susceptibility to subjectivity remains in the definition of resolution-shell ranges and limits, but the impact is minimal and hardly worth the effort. On the other hand, modifying the indicator to remove the shells so that the mean signal-to-noise ratio applied to all data above an appropriately adjusted limit value would extinguish that source of subjectivity.

When a cutoff is applied based on nominal resolution, then considerations similar to those for intensity-based cutoffs apply.

Copyright © 2012 International Union of Crystallography

If the objective is to estimate electron-density distributions, exclusion of large numbers of weak intensity terms beyond a limiting resolution is unlikely to be of any significant impact except, perhaps, a favourable one in computation time. If a sparse population of more intense reflections is also excluded, the effect may also be positive by reducing aberrations that might interfere with interpretation. If the objective is refinement, the benefit is less clear, except possibly in computation time. The cost for that may be the exclusion of a few more intense terms that provide positive guidance to refinement. Another positive effect of a sharply defined limiting resolution might be improved estimation of the spherical interference function where it is needed. It appears that, in practice, the consensus is that, in applying a cutoff, the exclusion of a few intense terms is at worst of negligible impact overall.

Because many imposed limits that exclude data are dependent on estimates of standard uncertainty, procedures for estimation of standard uncertainty need to be considered. As suggested earlier, a diffraction data set without a standard uncertainty for each intensity measurement is certainly the exception in current practice. The methods used by data-processing programs to estimate these standard uncertainties may be difficult, even impossible, to discern, but at core they must all be based on counting statistics. With the continuing trend toward data sets of high multiplicity (also referred to as redundancy), however, estimates of standard uncertainties are available from distributions of replicate measurements about means. Both of these estimates have value and express impacts of somewhat different sources of error. It would seem therefore that, where applicable, the best way to accommodate both would be to calculate weighted average values of individual intensities, where the weights are derived from the standard uncertainties from data processing, and then to estimate the standard uncertainties for the weighted averages by application of standard propagation-of-error procedures. This, finally, is the third area we highlight here that deserves focused discussion within the MX community with the objective of defining consensus best practices.

Little remains to be included in these introductory preliminaries. It should be understood that most of these indicators of quality may be cast in terms of either structure factors or intensities. We draw little attention to this as we define individual indicators, except that, where one of the two possibilities seems to dominate in usage, we tend to define that form alone. It is also true that many of these indicators have counterparts in which individual reflection terms are weighted. In current practice, these forms find little use in application to biological macromolecules and we ignore them here. Finally, many of these indicators may be expressed as fractions, that is, as numbers between one and zero. They are also often expressed in the percentage form, that is, as numbers between one hundred and zero. While we confine our use to the former here, we make no statement of preference.

The remainder of this chapter consists primarily of individual sections that reflect the various steps from the crystallographic data-collection experiment to the refinement of the final model. Relevant quality indicators and definitions are given for each of the steps and the most commonly used are collected in Table 2.2.11.1.

2.2.2. Quality indicators for diffraction data

Once useful crystalline samples have been obtained, the collection of X-ray diffraction data is the next (and the last) experimental step in a structure determination. Although the greatest care may be taken to collect data of as high quality as possible, there remain circumstances and influences that limit the quality of the data. Over time, many indicators have been defined to describe various aspects of diffraction data quality. The most important ones are discussed here.

Nominal resolution, $d_{\min}$. The resolution of a diffraction data set describes the extent of measurable data and is calculated by Bragg's law [equation (2.2.2.1)] based on the maximum Bragg angle 2θ included in the data set for a given data-collection wavelength λ. As discussed above, the nominal resolution is a limit set by the experimenters and is well known to be prone to subjective judgment. A number of suggestions have been made to reduce the subjectivity associated with this limit. One defines the limit as the resolution within which the intensities of a fraction of the unique reflections, for example 70%, are above a threshold, for example zero or three times their standard uncertainties. Another suggested limiting criterion, discussed further below, recommends that the nominal resolution be set as the midpoint of the resolution range of the shell at which the mean signal-to-noise ratio falls below 2.

$$d_{\min} = \lambda / 2 \sin \theta. \qquad (2.2.2.1)$$

True resolution, $d_{\rm true}$. The true resolution of a diffraction data set is defined as the minimum distance between two objects in a crystal that permits their images in the resultant electron-density map to be resolved. Often, $d_{\rm true}$ is approximated as $d_{\min}$.

To illustrate this crucial distance, represent two equivalent atoms by equal overlapping Gaussians. One might then consider that the distance between them that just permits distinguishing them as individual atoms might be the distance at which the electron-density value at the midpoint between the atoms drops to a value just below that at the positions of the atoms. For a normal distribution, this distance is 2σ, twice the standard deviation of the distribution.

Another perspective is provided by the realization that, when a Fourier synthesis is terminated at a resolution cutoff $d_{\min}$, successive spheres of negative and positive density of decreasing amplitude surround the maxima of positive density at atomic positions in that synthesis. It has been shown that the distance from the centre of such a maximum to the first zero is $0.715d_{\min}$ (James, 1948), which is a useful estimate of the limiting distance between distinguishable features in the electron-density map or $d_{\rm true}$. Similar estimates are used in other areas, notably for defining resolution in astronomy. A more recent re-evaluation suggests that a limit of $0.917d_{\min}$ is a better value, especially when the effects of form factors and atomic displacement parameters are considered (Stenkamp & Jensen, 1984). Add to that the effects of errors in experimental amplitudes and derived phases, and the approximation of $d_{\rm true}$ as $d_{\min}$ seems quite reasonable.

Optical resolution, $d_{\rm opt}$. The optical resolution $d_{\rm opt}$ is calculated from the standard deviation of a Gaussian fitted to the origin peak of the Patterson function ($\sigma_{\rm Patt}$) of the diffraction data set and the standard deviation of another Gaussian fitted to the origin peak of the spherical interference function ($\sigma_{\rm sph}$). This definition is based on Vaguine *et al.* (1999) and is implemented in the program *SFCHECK*. The optical resolution is intended to account for uncertainties in the data, atomic displacement factors, effects of crystal quality and series-termination effects by means of a propagation-of-error-like approach (Blundell & Johnson, 1976; Vaguine *et al.*, 1999). It has been suggested that $d_{\rm opt}$ is a better approximation of $d_{\rm true}$ than

d_{min} (Weiss, 2001).

$$d_{\mathrm{opt}} = [2(\sigma^2_{\mathrm{Patt}} + \sigma^2_{\mathrm{sph}})]^{1/2}. \qquad (2.2.2.2)$$

Completeness, *C*. The completeness *C* of a diffraction data set is defined as the fraction of the unique reflections in a given space group to a given nominal resolution d_{min} that have been measured at least once during data collection. *C* may be given assuming that Friedel symmetry is either applied or not. In the latter case, *C* is also referred to as anomalous completeness. In the program *SCALA* (Evans, 2006), the anomalous completeness is defined based on acentric reflections only.

Effective resolution, d_{eff}. Since any missing reflection of a data set leads to a deterioration of the model parameters (Hirshfeld & Rabinovich, 1973; Arnberg *et al.*, 1979), an effective resolution may be defined based on the nominal resolution d_{min} and the cube root of the completeness *C* of the data set.

$$d_{\mathrm{eff}} = d_{\mathrm{min}} C^{-1/3}. \qquad (2.2.2.3)$$

Multiplicity (or redundancy), *N*. The multiplicity or redundancy *N* of a diffraction data set defines on average how many times a reflection *hkl* has been observed during the data-collection experiment including symmetry mates and replicate measurements. *N* may be given assuming that Friedel symmetry is either applied or not.

Merging *R* factor, R_{merge}. The merging *R* factor of a diffraction data set describes the spread of the individual intensity measurements I_i of a reflection *hkl* around the mean intensity $\langle I(hkl)\rangle$ of this reflection. Sometimes R_{merge} is also referred to as *R* factor (observed) (in the program *XDS*; Kabsch, 1988, 1993, 2010; Chapter 11.6), as R_{sym} or as R_{linear} (in the program *SCALEPACK*; Otwinowski & Minor, 1997). In fractional form, this is

$$R_{\mathrm{merge}} = \sum_{hkl}\sum_{i} |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_{i} I_i(hkl), \qquad (2.2.2.4)$$

where $\langle I(hkl)\rangle$ is the mean of the several individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*. The sums $\sum_{hkl}$ and $\sum_i$ run over all observed unique reflections *hkl* and over all individual observations *i* of a given reflection *hkl*. It should be noted that alternative definitions of R_{merge} exist. In one, $I_i(hkl)$ in the denominator is replaced by $\langle I(hkl)\rangle$, thereby producing an expression that is formally equivalent to the one above. In another, $I_i(hkl)$ in the denominator is replaced by $|I_i(hkl)|$ with the suggestion that the denominator is thereby prevented from becoming negative or zero, even in the case of many negative-intensity observations. One should note, however, the counter-intuitive side effect: artificial damping of R_{merge} values, that is, reducing expected higher R_{merge} values of data sets with more weak reflections.

The usefulness of R_{merge} as a quality indicator for diffraction data is limited because it is dependent on the multiplicity of a data set (Diederichs & Karplus, 1997*a*,*b*; Weiss & Hilgenfeld, 1997; Weiss, 2001). The higher the multiplicity of a data set, the higher its R_{merge} will be, although, based on statistics, the better determined the averaged intensity values should be. Despite these shortcomings, R_{merge} is still widely used today.

Redundancy-independent merging *R* factor, $R_{\mathrm{r.i.m.}}$ or R_{meas}. The redundancy-independent merging *R* factor $R_{\mathrm{r.i.m.}}$ or R_{meas} describes the precision of the individual intensity measurements I_i, independent of how often a given reflection has been measured. Because of its independence of the redundancy (hence its name), it has been proposed that $R_{\mathrm{r.i.m.}}$ or R_{meas} should be used as a substitute for the conventional R_{merge} (Diederichs & Karplus, 1997*a*,*b*; Weiss & Hilgenfeld, 1997; Weiss, 2001). In fractional form, this is

$$R_{\mathrm{r.i.m.}} = \sum_{hkl}\{N(hkl)/[N(hkl)-1]\}^{1/2} \times \sum_{i} |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_{i} I_i(hkl), \qquad (2.2.2.5)$$

where $\langle I(hkl)\rangle$ is the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*. As for R_{merge}, the sums $\sum_{hkl}$ and $\sum_i$ run over all observed unique reflections *hkl* and over all individual observations *i* of a given reflection *hkl*.

Precision-indicating merging *R* factor, $R_{\mathrm{p.i.m.}}$. The precision-indicating merging *R* factor $R_{\mathrm{p.i.m.}}$ describes the precision of the averaged intensity measurements $\langle I(hkl)\rangle$ (Weiss, 2001). In fractional form, this is

$$R_{\mathrm{p.i.m.}} = \sum_{hkl}\{1/[N(hkl)-1]\}^{1/2} \times \sum_{i} |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_{i} I_i(hkl), \qquad (2.2.2.6)$$

where $\langle I(hkl)\rangle$ is the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*. As with R_{merge} and $R_{\mathrm{r.i.m.}}$ or R_{meas}, the sums $\sum_{hkl}$ and $\sum_i$ run over all observed unique reflections *hkl* and over all individual observations *i* of a given reflection *hkl*.

***R* factor of merged intensities or amplitudes, $R_{\mathrm{mrgd}\text{-}I}$ and $R_{\mathrm{mrgd}\text{-}F}$.** An alternative precision-indicating merging *R* factor, called R_{mrgd}, is defined as the *R* factor between two or more data sets or between two subsets of a data set created by randomly apportioning the individual intensity measurements between the two subsets (Diederichs & Karplus, 1997*a*,*b*). R_{mrgd} can be calculated for intensities ($R_{\mathrm{mrgd}\text{-}I}$) or structure-factor amplitudes ($R_{\mathrm{mrgd}\text{-}F}$). The latter quantity was suggested to present a lower limit for the crystallographic *R* factor of a model against the observed data (Diederichs & Karplus, 1997*a*,*b*). In fractional form

$$R_{\mathrm{mrgd}\text{-}I} = 2\sum_{hkl} |\langle I_1(hkl)\rangle - \langle I_2(hkl)\rangle| / \sum_{hkl}\langle I_1(hkl)\rangle + \langle I_2(hkl)\rangle, \qquad (2.2.2.7)$$

where $\langle I_1(hkl)\rangle$ and $\langle I_2(hkl)\rangle$ are the mean intensity values for the individual observations of the reflections *hkl*, which have been partitioned into the two subsets 1 and 2. The sums $\sum_{hkl}$ run over all observed unique reflections. $R_{\mathrm{mrgd}\text{-}I}$ is related to $R_{\mathrm{p.i.m.}}$ by a constant factor ($R_{\mathrm{mrgd}\text{-}I} = 2^{1/2} R_{\mathrm{p.i.m.}}$).

$R_{\mathrm{mrgd}\text{-}F}$ is defined analogously to $R_{\mathrm{mrgd}\text{-}I}$ (Diederichs & Karplus, 1997*a*,*b*). In the equation, only the intensities are replaced by structure-factor amplitudes.

$$R_{\mathrm{mrgd}\text{-}F} = 2\sum_{hkl} |\langle F_1(hkl)\rangle - \langle F_2(hkl)\rangle| / \sum_{hkl}\langle F_1(hkl)\rangle + \langle F_2(hkl)\rangle. \qquad (2.2.2.8)$$

In order to cope with negative-intensity observations, pseudo-amplitudes had to be introduced just for the purpose of calculating $R_{\mathrm{mrgd}\text{-}F}$ ($F = I^{1/2}$ if $I \geq 0$ and $F = -|I|^{1/2}$ if $I < 0$).

Note. The approach of comparing randomly partitioned subsets of a given data set is used for a variety of quality indicators. While there is potential for variation in these indicators from one partitioning of the data set to another, an average of several random partitionings should be expected to give a useful estimate. There is also potential for subjectivity, but the principal value of these indicators is to assist the experimenter in proper analysis and they are less often applied to compare experiments from different laboratories and are seldom published.

Pooled coefficient of variation, PCV. The pooled coefficient of variation PCV is the ratio of the sum of the standard deviations to the sum of the reflection intensities (Diederichs & Karplus, 1997*a*,*b*). PCV is related to $R_{\rm meas}$ or $R_{\rm r.i.m.}$ by the factor $(\pi/2)^{1/2}$. In fractional form, this is

$$\mathrm{PCV} = \frac{\sum_{hkl}\left(\{1/[N(hkl)-1]\}^{1/2}\sum_i |I_i(hkl)-\langle I(hkl)\rangle|^2\right)^{1/2}}{\sum_{hkl}\sum_i I_i(hkl)}, \tag{2.2.2.9}$$

where $\langle I(hkl)\rangle$ is the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of reflection hkl.

Mean signal-to-noise ratio, ⟨*I*/σ(*I*)⟩. The signal-to-noise ratio $I_i/\sigma(I_i)$ of an individual intensity measurement describes the statistical significance of a measured intensity. As a measure of the overall quality of a data set, the mean signal-to-noise ratio for all reflections is useful as an indication of the robustness of the data, that is, the average intensity as a multiple of the standard uncertainty. In addition, as mentioned above, the mean signal-to-noise ratio for all reflections within the outer resolution shell can be used to define the nominal resolution of a data set. For the data set as a whole or for a resolution shell of that data set, the mean signal-to-noise ratio, $\langle I/\sigma(I)\rangle$, is the sum of the signal-to-noise ratios of all individual reflections hkl within resolution limits divided by the number of individual reflections hkl within those resolution limits.

In principle there are two ways to define a mean signal-to-noise ratio of a data set (or a given resolution shell). The two ways yield different quantities, although, unfortunately, they are both called the mean signal-to-noise ratio. They differ in the manner in which mean signal-to-noise ratios of individual reflections hkl are calculated.

(i) $\langle I(hkl)\rangle/\sigma[I(hkl)]$. The mean signal-to-noise ratio of individual reflections hkl may be calculated as the ratio of the mean intensity $\langle I(hkl)\rangle$ and the r.m.s. scatter of $I_i(hkl)$ about that mean. This is a measure of the average significance of individual observations, but it does not take into account the multiplicity or redundancy of the measurements.

$$\begin{aligned}&\langle I(hkl)\rangle/\sigma[I(hkl)]\\ &\quad= \langle I(hkl)\rangle/[(1/N)\sum_i |I_i(hkl)-\langle I(hkl)\rangle|^2]^{1/2}.\end{aligned} \tag{2.2.2.10}$$

In the program *SCALA* (Evans, 2006), this value is reported as I/sigma.

(ii) $\langle I(hkl)\rangle/\sigma\langle I(hkl)\rangle$. $\langle I(hkl)\rangle$ is the average over all observations of the reflection hkl, and is sometimes weighted. $\sigma\langle I(hkl)\rangle$ is the propagation-of-error combination of standard uncertainties assigned at data processing for the individual measurements $I_i(hkl)$, that is, a modification of equation (2.2.2.10) in which the term $|I_i(hkl)-\langle I(hkl)\rangle|$ in the denominator is replaced by $\sigma_i(hkl)$, the experimental standard uncertainty for the measurement $I_i(hkl)$.

$$\langle I(hkl)\rangle/\sigma\langle I(hkl)\rangle = \langle I(hkl)\rangle/[(1/N)\sum_i \sigma_i(hkl)^2]^{1/2}. \tag{2.2.2.11}$$

An error model[1] is often applied in the denominator here to scale to the r.m.s. scatter in (i) above. In the program *SCALA* (Evans, 2006), this value is reported as Mn(I)/sd.

Both methods of defining the mean signal-to-noise ratio for the reflection hkl have merit. As suggested for individual intensities in Section 2.2.1, perhaps the best approach would be to calculate weighted averages and weighted standard uncertainties of the $I(hkl)$ where weights are the experimental standard uncertainties $\sigma_i(hkl)$ for individual measurements $I_i(hkl)$.

Highest possible signal-to-noise ratio, *I*/σ(*I*)$_{\rm asymptotic}$. A relatively recent addition to the collection of diffraction-data quality indicators is the highest possible signal-to-noise ratio of a data set $I/\sigma(I)_{\rm asymptotic}$ or ISa (Diederichs, 2010). ISa is calculated from the parameters of the error model used for inflating the standard deviations of the reflections with an intensity-dependent term.[1] Since ISa is practically independent of counting statistics, it was suggested to be a good measure of instrument errors manifesting themselves in the data set, provided the crystal is close to ideal and radiation damage is negligible. Data sets with ISa values of 25 or greater are considered to be very good and amenable to straightforward structure determination, while data sets exhibiting ISa values of 15 or less are considered marginal at best. The calculation of ISa is implemented in *XDS* versions of December 2009 or later (Kabsch, 2010).

Anomalous *R* factor, *R*$_{\rm anom}$. The anomalous R factor $R_{\rm anom}$ describes the sum of the differences in intensities of Friedel-related reflections (hkl) and $(\bar{h}\bar{k}\bar{l})$ relative to the sum of their mean intensities. In fractional form, this is

$$R_{\rm anom} = \sum_{hkl} |I(hkl) - I(\bar{h}\bar{k}\bar{l})|/\sum_{hkl}\langle I(hkl)\rangle, \tag{2.2.2.12}$$

where, in this case, $\langle I(hkl)\rangle$ is the mean intensity of the Friedel mates of the reflections hkl, or $\frac{1}{2}[I(hkl)+I(\bar{h}\bar{k}\bar{l})]$. Here, the sums $\sum_{hkl}$ run over all unique reflections with one of the indices, typically h, greater than zero ($h > 0$) for which both Friedel mates have been observed at least once.

The ratio of $R_{\rm anom}$ to $R_{\rm p.i.m.}$ has been proposed as a possible indicator for the strength of the anomalous signal (Panjikar & Tucker, 2002).

Anomalous correlation coefficient, CC$_{\rm anom}$. The anomalous correlation coefficient CC$_{\rm anom}$ quantifies the linear dependence of observed anomalous differences in two diffraction data sets. These can be data sets, for example, collected at two different wavelengths in a MAD experiment. In cases where only one data set is available, two randomly partitioned half data sets can be created for comparison.

Note. The correlation coefficient referred to here and elsewhere in this chapter is invariably the Pearson linear correlation coefficient (Rodgers & Nicewander, 1988):

$$\mathrm{CC} = \sum(x-\langle x\rangle)(y-\langle y\rangle)\Big/\Big[\sum(x-\langle x\rangle)^2\sum(y-\langle y\rangle)^2\Big]^{1/2}, \tag{2.2.2.13}$$

with x and y being, in this case, the anomalous differences $[I(hkl) - I(\bar{h}\bar{k}\bar{l})]$ or $[F(hkl) - F(\bar{h}\bar{k}\bar{l})]$ in the two data sets, $\langle x\rangle$ and $\langle y\rangle$ are their averages, and the summations are over all reflections

---

[1] Typically, the experimental standard uncertainties are modified by various correction factors in order to account for errors in the data that arise from other sources than counting statistics. The inflation factors are applied in various ways in different programs. In *SCALA* (Collaborative Computational Project, Number 4, 1994; Evans, 2006) they are called SDFAC, SDB and SDADD. In *SCALEPACK* (Otwinowski & Minor, 1997), they are called ERROR SCALE FACTOR and ESTIMATED ERROR. In *D*TREK* (Pflugrath, 1999), they are called E$_{\rm mul}$ and E$_{\rm add}$. In *XDS* (Kabsch, 1988, 1993, 2010; Chapter 11.6) they are called a and b.

hkl for which observations exist in both data sets across the entire resolution range or within a particular resolution shell. $\mathrm{CC}_{\mathrm{anom}}$ is a reliable indicator of the strength of the anomalous signal. Values above 0.30 are considered good.

R.m.s. correlation ratio. This is another statistic based on randomly partitioned data sets, which is calculated by the program *SCALA* (Evans, 2006; Collaborative Computational Project, Number 4, 1994). It is an analysis of the scatterplot of $[I_1(hkl) - I_1(\bar{h}\bar{k}\bar{l})]$ *versus* $[I_2(hkl) - I_2(\bar{h}\bar{k}\bar{l})]$, where the subscripts 1 and 2 identify the two half data sets. The analysis assumes that the correlation is ideally 1.0. The r.m.s. correlation ratio is defined as the ratio of the r.m.s. widths of the scatterplot distribution along the diagonal and perpendicular to the diagonal. This statistic seems to be more robust than $\mathrm{CC}_{\mathrm{anom}}$ to the presence of outliers. It cannot, however, be applied to analysing the correlations between different data sets.

Mean anomalous signal-to-noise ratio, $\langle d''/\sigma(d'')\rangle$. The anomalous signal-to-noise ratio of an individual reflection measurement $d''(hkl)/\sigma[d''(hkl)]$ is defined as the ratio of the observed anomalous intensity difference $d'' = |I(hkl) - I(\bar{h}\bar{k}\bar{l})|$ and the corresponding estimated standard uncertainty $\sigma(d'')$ in the measurement of this anomalous difference. The average of the anomalous signal-to-noise ratios for all reflections within a certain resolution range is used as an indicator of utility for phasing. A value of $(2/\pi)^{1/2} \simeq 0.8$ for mean $d''/\sigma(d'')$ of a resolution shell, for example, is taken to indicate that no anomalous signal is present (G. Sheldrick & G. Bunkoczi, personal communication).

Decay *R* factor, R_d. The decay *R* factor R_d is defined as a pairwise *R* factor based on the intensities of symmetry-related reflections occurring on different diffraction images (Diederichs, 2006). An increase in R_d as a function of difference in image-collection times is a good indicator of radiation damage occurring during data collection. In fractional form, this is

$$R_d = 2\sum_{hkl}\sum_{m-n}|I_m(hkl) - I_n(hkl)|/\sum_{hkl}\sum_{i}[I_m(hkl) + I_n(hkl)], \tag{2.2.2.14}$$

where $I_m(hkl)$ and $I_n(hkl)$ are the intensities of the reflection *hkl* occurring on images *m* and *n*. The only program in which this is currently implemented is *XDSSTAT*.

Wilson-plot *B* factor, B_{Wilson}. A Wilson plot (Wilson, 1949) is a plot for a contiguous series of resolution shells of the logarithm of the mean intensity in a given resolution shell divided by the sum of the squared atomic form factors for all atoms in the unit cell evaluated at the mean of the resolution limits of the shell. From a least-squares fit of a straight line to the linear part of the Wilson plot, the *B* factor B_{Wilson} can be derived. Typically, data of lower than 4.5 Å resolution are excluded from the fit. The more meaningful determinations of B_{Wilson} come from Wilson plots that are linear all the way to the nominal resolution $d_{\min}$ and minimize the occurrence of spikes due to ice rings.

$$\ln(\langle I_{\mathrm{obs}}\rangle/\sum_i f_i^2) = -\ln K_{\mathrm{Wilson}} - 2B_{\mathrm{Wilson}}/d^2, \tag{2.2.2.15}$$

where $\langle I_{\mathrm{obs}}(hkl)\rangle$ is the mean over the intensities of all observed reflections *hkl* in a given resolution shell. The sum $\sum_i$ runs over all atoms in the structure. The parameter *d* is the midpoint of the resolution shell over which I_{obs} has been averaged. K_{Wilson} is an absolute scale factor.

2.2.3. Comparing different diffraction data sets

In MX, there are many instances when two or more diffraction data sets have to be compared to each other. An important decision which has to be made is, for instance, whether data sets collected from different crystals are isomorphous enough so that they can be merged together. Another task is the comparison of a native data set and a heavy-atom derivative. The question here is how much of the observed difference is due to non-isomorphism and how much is due to isomorphous replacement.

Scaling *R* factor, R_{scale}. The scaling *R* factor R_{scale} between two data sets is defined as the difference in the structure-factor amplitudes of one data set relative to the structure-factor amplitudes of the other data set after the two data sets have been scaled to each other.

$$R_{\mathrm{scale}} = \sum_{hkl}|F_1 - F_2|/\sum_{hkl}|F_1|. \tag{2.2.3.1}$$

However, the above formula is asymmetric with respect to data sets 1 and 2. An alternative symmetric formula is

$$R_{\mathrm{scale}} = 2\sum_{hkl}|F_1 - F_2|/\sum_{hkl}(F_1 + F_2). \tag{2.2.3.2}$$

R_{scale} is also often given based on intensities rather than structure-factor amplitudes.

Gradient from a normal probability analysis. For each reflection *hkl* in two data sets that have been scaled to each other, the quantity $\Delta(\mathrm{real})$ is computed and compared with its expectation value $\Delta(\mathrm{expected})$ assuming a normal distribution of errors (Howell & Smith, 1992). If $\Delta(\mathrm{real})$ is plotted against $\Delta(\mathrm{expected})$, random differences between the two data sets give rise to a slope of 1.0. Significant differences result in slopes significantly larger than 1.0. Such an analysis is implemented in the program *SCALEIT* (Collaborative Computational Project, Number 4, 1994).

$$\Delta(\mathrm{real}) = (F_1 - F_2)/[\sigma(F_1)^2 + \sigma(F_2)^2]^{1/2}. \tag{2.2.3.3}$$

2.2.4. Quality indicators for substructure determination

The first step of a crystal structure determination after processing and analysing the diffraction data is to use the observed isomorphous and/or anomalous differences in order to determine the isomorphous or anomalous substructure. Since such substructures typically consist of rather few atoms, which are relatively far apart, direct-methods approaches have been adopted from small-molecule crystallography and have largely superseded pure Patterson-based methods. The most widely used computer programs for substructure determination are *SHELXD* (Schneider & Sheldrick, 2002; Sheldrick, 2008), *SnB* (Weeks *et al.*, 1993, 1994) and *HySS* (Grosse-Kunstleve & Adams, 2003).

Correlation coefficient CC(all). CC(all) measures the correlation of the *E* values (analogous to normalized structure factors; Hauptman & Karle, 1953) derived from the observed isomorphous or anomalous differences and those calculated from the substructure model. CC(all) is another application of the Pearson linear correlation coefficient [see equation (2.2.2.13)] and is calculated for all reflections. In the equation, *x* and *y* are the *E* values derived from the observed and calculated differences. CC(all) is typically used to assess the successful determination of the isomorphous or anomalous substructure in a macromolecule. A value of CC(all) ≥ 0.3 often indicates that the substructure has been correctly identified (Schneider & Sheldrick, 2002).

Correlation coefficient CC(weak). Like CC(all), CC(weak) is the Pearson linear correlation coefficient [see equation (2.2.2.13)] between the E values (analogous to normalized structure factors as above) derived from the observed isomorphous or anomalous differences and those calculated from the substructure model. In contrast to CC(all), however, CC(weak) is calculated for the weak reflections only. As above, x and y are the E values derived from the observed and calculated differences. The E-value cutoff for defining a reflection as weak can be chosen by the user, but a typical threshold value is 1.5, although lower values may be required for low-resolution data (Schneider & Sheldrick, 2002). A value of CC(weak) ≥ 0.15 often indicates that the substructure has been correctly identified (Sheldrick, 2010).

The minimal function, $R(\varphi)$. The minimal function $R(\varphi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle).

$$R(\varphi) = \sum_{H,K} A_{HK}\left\{\cos\varphi_{HK} - [I_1(A_{HK})/I_0(A_{HK})]\right\}^2 \Big/ \sum_{H,K} A_{HK}, \tag{2.2.4.1}$$

where $A_{HK} = (2/N^{1/2})|E_H E_K E_{H+K}|$ with N being the number of atoms in the corresponding primitive unit cell. The minimal function is the function minimized in the program *Shake&Bake*, abbreviated as *SnB* (Weeks *et al.*, 1993, 1994).

2.2.5. Quality indicators for phase determination

Once the isomorphous or anomalous substructure has been successfully determined, it can be used as reference point for the calculation of phases. The quality of the resulting phases is dependent on the strength of the isomorphous or anomalous signal and the completeness and correctness of the isomorphous or anomalous substructure.

Cullis *R* factor, $R_{\rm Cullis}$. The Cullis R factor (Cullis *et al.*, 1961) for phase determination by isomorphous replacement is defined as the ratio between the lack-of-closure error $\varepsilon(\varphi_{\rm P})$ [equation (2.2.5.1*a*) below] and the isomorphous difference $|F_{\rm PH} - F_{\rm P}|$. It is the most useful signal for a usable heavy-atom derivative. Values < 0.6 for centrosymmetric data are excellent, while values < 0.9 are still usable.

$$\varepsilon(\varphi_{\rm P}) = \big|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\big|, \tag{2.2.5.1a}$$

$$R_{\rm Cullis} = \sum_{hkl} \big|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\big| \Big/ \sum_{hkl} |F_{\rm PH} - F_{\rm P}|. \tag{2.2.5.1b}$$

Anomalous Cullis *R* factor, $R_{\rm Cullis,ano}$. The Cullis R factor for phase determination by anomalous dispersion is defined as the ratio between the lack-of-closure error and the observed anomalous difference $|F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})|$. The lack-of-closure error in the anomalous-dispersion case is the difference between the observed anomalous difference and the calculated anomalous difference $2F_{\rm H}\sin\alpha_{\rm P}$, where $\alpha_{\rm P}$ is the protein phase. A value of $R_{\rm Cullis,ano} < 1.0$ suggests that a contribution to the phasing from the anomalous data is likely (*MLPHARE* program documentation; Collaborative Computational Project, No. 4, 1994).

$$R_{\rm Cullis,ano} = \frac{\sum_{hkl} \big||F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})| - |2F_{\rm H}\sin\alpha_{\rm P}|\big|}{\sum_{hkl} |F_{\rm PH}(hkl) - F_{\rm PH}(\bar{h}\bar{k}\bar{l})|}. \tag{2.2.5.2}$$

Phasing power, $\mathbf{PP}_{\rm iso}$. The isomorphous phasing power $\mathrm{PP}_{\rm iso}$ for phase determination by isomorphous replacement is defined for a particular pair of native and heavy-atom-derivative data sets as the ratio of $|F_{\rm H}|$ and $\varepsilon(\varphi_{\rm P})$, where $|F_{\rm H}|$ is the calculated amplitude of the heavy-atom structure factor and $\varepsilon(\varphi_{\rm P})$ is the lack-of-closure error [equation (2.2.5.1*a*)]. $\mathbf{F}_{\rm P} + \mathbf{F}_{\rm H}$ is a vector sum of the calculated structure factor $\mathbf{F}_{\rm H}$ and the structure factor $\mathbf{F}_{\rm P}$.

$$\mathrm{PP}_{\rm iso} = \sum_{hkl} |F_{\rm H}| \Big/ \sum_{hkl} \big|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\big|. \tag{2.2.5.3}$$

There is another, slightly different, definition of $\mathrm{PP}_{\rm iso}$, which is implemented in the program *SOLVE*. Here, $\mathrm{PP}_{\rm iso}$ is defined as the ratio of the r.m.s. of the $|F_{\rm H}|$ values and the r.m.s. of the lack-of-closure errors $\varepsilon(\varphi_{\rm P})$. For each reflection hkl, a weighted average of $\varepsilon(\varphi_{\rm P})$ is calculated by integrating $\varepsilon(\varphi_{\rm P})$ over the whole phase circle.

$$\mathrm{PP}_{\rm iso} = \Big(\sum_{hkl} |F_{\rm H}|^2\Big)^{1/2} \Big/ \Big(\sum_{hkl} \langle\big|F_{\rm PH} - |F_{\rm P} + F_{\rm H}|\big|\rangle^2\Big)^{1/2}. \tag{2.2.5.4}$$

Note. Owing to the cancelling out of the factor $N^{1/2}$ in the numerator and denominator, the latter $\mathrm{PP}_{\rm iso}$ formula does not appear as a ratio of r.m.s. values at first sight.

Anomalous phasing power, $\mathbf{PP}_{\rm ano}$. The anomalous phasing power $\mathrm{PP}_{\rm ano}$ for phase determination by anomalous-dispersion methods is defined as the ratio of the sum of calculated anomalous differences $d''_{\rm calc}$ and the sum of estimated standard uncertainties $\sigma(d''_{\rm obs})$ in the measurement of these anomalous differences:

$$\mathrm{PP}_{\rm ano} = \sum_{hkl} d''_{\rm calc} \Big/ \sum_{hkl} \sigma(d''_{\rm obs}). \tag{2.2.5.5}$$

As with $\mathrm{PP}_{\rm iso}$ (see above), the program *SOLVE* uses a slightly different definition of $\mathrm{PP}_{\rm ano}$. Here, the anomalous phasing power is defined as the ratio of the r.m.s. of the $d''_{\rm calc}$ values and the r.m.s. of $\sigma(d''_{\rm obs})$. For this, a weighted average of $d''_{\rm calc}$ is computed by integrating over the whole phase circle for each reflection.

$$\mathrm{PP}_{\rm ano} = \Big(\sum_{hkl} \langle d''_{\rm calc}\rangle^2\Big)^{1/2} \Big/ \Big[\sum_{hkl} \sigma(d''_{\rm obs})^2\Big]^{1/2}. \tag{2.2.5.6}$$

Note. As above in the $\mathrm{PP}_{\rm iso}$ formula, the factors $N^{1/2}$ in the numerator and denominator cancel out.

Figure of merit (f.o.m.), *m*. The figure of merit m is a term used in a number of contexts in X-ray crystallography. In its most common use, it is defined as the weight applied to an individual structure-factor amplitude that, in conjunction with its best phase, gives rise, in a Fourier synthesis, to the electron-density map with the minimum level of noise (Blow & Crick, 1959). Typically, m is given as an average value over all reflections in the data set or in a given resolution shell.

$$m = \int P(\alpha)\exp(i\alpha)\,{\rm d}\alpha \Big/ \int P(\alpha)\,{\rm d}\alpha = \langle\cos(\Delta\alpha)\rangle, \tag{2.2.5.7}$$

where $P(\alpha)$ is the probability of the phase α, initial or refined, being the best phase and $\Delta\alpha = \alpha_{\rm best} - \alpha$ is the error in the phase angle at α. The integration is from 0 to 2π and values for m range from 0 to 1.

2.2.6. Quality indicators for density modification and phase improvement

After determination of initial phases, a first electron-density map can be computed. It is expected that this map will contain significant errors and improbable features. Additional information, such as the flatness of the electron density in the solvent region or the similarity of electron-density regions of two or more identical molecules in the asymmetric unit, can be exploited to modify the electron density and hence improve the phases.

Figure of merit (f.o.m.), *m*. The figure of merit m [see equation (2.2.5.7)] is also used to judge the degree of improvement in the phase values. Again, in its most common use, it is defined as the weight applied to an individual structure-factor amplitude that, in conjunction with its best phase, gives rise, in a Fourier synthesis, to the electron-density map with the minimum level of noise. Typically, m is given as an average value over all reflections in the data set or in a given resolution shell.

Density-modification (DM) free *R* factor. The DM free R factor is defined in the same way as the refinement free R factor $R_{\rm free}$ [see equation (2.2.8.1) and the following paragraph describing $R_{\rm free}$]. It describes the disagreement between the observed structure-factor amplitudes of a certain set of reflections and the calculated amplitudes after density modification. It is a weak indicator used in the program *DM* (Cowtan, 1999) and is helpful mainly in identifying the correct enantiomorph.

Density-modification (DM) real-space free residuals. The DM real-space free residuals are two numbers (real-space free residual for the solvent area and real-space free residual for the protein area) which are calculated by omitting two small spheres of protein and solvent from the density-modification process. The real-space free residual for the solvent area indicates how flat the solvent is in a non-flattened region of solvent and the real-space free residual for the protein area indicates how well the electron density in a non-histogram-matched region of protein obeys the expected histogram. The two residuals can also be combined into the combined real-space free residual by weighted summation, where the weight is defined by the solvent content. The DM real-space free residuals have some value in determining when to stop a density-modification calculation, once no more progress is being made, but limited value otherwise.

Contrast, *c*. The contrast c between the r.m.s. electron density in the solvent region and the r.m.s. electron density in the macromolecular region can serve as an indication of the presence of a clearly-defined solvent boundary in the electron-density map. It is defined as the standard deviation of the local r.m.s. electron-density values over the entire asymmetric unit (Terwilliger & Berendzen, 1999; Sheldrick, 2002; Terwilliger *et al.*, 2009). The electron-density values are first squared (optionally after truncating very high and very low density values) and then smoothed using a moving local sphere typically with a radius of 6 Å. Local r.m.s. values are then calculated. The contrast c is now defined as the standard deviation σ of the local r.m.s. density values multiplied by a factor to normalize for the fraction of solvent sc in the crystal:

$$c = [(1 - sc)/sc]^{1/2}\sigma. \qquad (2.2.6.1)$$

Skewness of electron density, *S*. A high value for the skewness S of the electron density in an electron-density map indicates the presence of local electron-density maxima with high positive density values. The skewness is defined as the third moment of the electron density:

$$S = \langle\rho^3\rangle/\langle\rho^2\rangle^{3/2}. \qquad (2.2.6.2)$$

In order to compute the mean values of ρ^3 and ρ^2, all density grid points in the asymmetric unit of the electron-density map are taken into account (Terwilliger *et al.*, 2009).

Overlap of NCS-related density, $O_{\rm NCS}$. The presence of correlated electron density at noncrystallographic symmetry (NCS)-related regions in a map can be used as an indicator for the quality of the electron-density map (Cowtan & Main, 1998; Vellieux *et al.*, 1995; Terwilliger *et al.*, 2009). The overlap ($O_{\rm NCS}$) between density values at NCS-related locations is also often used to evaluate the presence of local symmetry or non-space-group symmetry:

$$O_{\rm NCS} = \langle\rho_i\rho_j\rangle. \qquad (2.2.6.3)$$

ρ_i and ρ_j are the normalized electron-density values in the NCS-related regions in the asymmetric unit. The average is calculated over the whole region where NCS is present. This region may be defined as the region where overlap values are 0.3 or greater, or by a mask. If there are more than two NCS groups, the average is taken over all NCS pairs.

***R* factor, $R_{\rm DENMOD}$, and phase correlation, $m_{\rm DENMOD}$, from statistical density modification.** The amplitudes and phases of structure factors calculated using statistical density modification can be compared with the observed amplitudes and experimental phases (Cowtan & Main, 1996; Terwilliger, 2001; Terwilliger *et al.*, 2009). These comparisons yield an R value ($R_{\rm DENMOD}$) for the amplitudes and a mean cosine of the phase difference ($m_{\rm DENMOD}$) for the phases.

$$R_{\rm DENMOD} = \sum_{hkl} \big||F_{\rm obs}| - |F_{\rm DENMOD}|\big| / \sum_{hkl} |F_{\rm obs}|, \qquad (2.2.6.4)$$

$$m_{\rm DENMOD} = (1/N) \sum_{hkl} \cos(\alpha_{\rm obs} - \alpha_{\rm DENMOD}). \qquad (2.2.6.5)$$

Correlation coefficient CC of chain trace against native data. The quality of density modification in *SHELXE* (Sheldrick, 2002, 2010) can be assessed by computing a Pearson linear correlation coefficient [see equation (2.2.2.13)] of the calculated structure-factor amplitudes for a chain trace against the native structure-factor amplitudes. If the poly-Ala trace yields a CC value higher than 0.25 and if the mean chain length of the trace is >10, the solution is almost always correct. This criterion is currently implemented in the program *ARCIMBOLDO* (Rodriguez *et al.*, 2009).

2.2.7. Quality indicators for molecular replacement

In a case where a known structure is assumed to be similar to that of a target molecule (structural similarity is typically inferred by the degree of sequence similarity), the known structure, also termed the search model, can be used to determine the structure of the target molecule. The approach, termed molecular replacement, was first described in 1962 (Rossmann & Blow, 1962). Nowadays, about two thirds of all newly determined structures are determined by molecular replacement (Long *et al.*, 2008).

Rotation function, RF. The rotation function RF is a measure of the overlap or the agreement of the stationary Patterson function $P2$ calculated from the observed data and the rotated Patterson function $P1$ from the search model.

$$\mathrm{RF} = \int_r P2\, \underline{R}\, P1 \,\mathrm{d}r. \qquad (2.2.7.1)$$

In the equation for RF, $\underline{\mathrm{R}}$ is the rotation operator. The integration is performed between a minimum value and a maximum value for the radius r. These values are chosen according to the size of the search model, with the aim of including as many intramolecular Patterson peaks (self vectors) as possible and to exclude as many intermolecular Patterson peaks (cross vectors) as possible. The ratio of the height of a peak in the RF to the background level is used as an indicator of how likely it is that this peak describes the orientation of a molecule in the target structure.

Translation function, TF. There are numerous ways of defining a translation function TF, making it impractical to discuss quality indicators here. For a thorough treatment of translation-function applications, the reader is referred to Chapter 2.3 of *International Tables for Crystallography* Volume B and Chapter 13.3 of the present volume.

Log-likelihood gain, LLG. In likelihood-based molecular replacement (McCoy *et al.*, 2007), potential molecular-replacement solutions are evaluated using likelihood, which is defined as the probability P that the observed diffraction data would have been measured if the orientation (and, usually, position) of the model were correct (Read, 2001). The score is reported in terms of the log-likelihood gain (LLG), which is defined as the logarithm of the likelihood score for the model $p(F_{\rm obs}; \text{model})$ minus the logarithm of the likelihood score for a random-atom Wilson distribution $p_{\rm Wilson}(F_{\rm obs})$. The LLG measures how much better the data can be predicted from the molecular-replacement model than from a collection of random atoms.

$$\mathrm{LLG} = \sum_{hkl} \ln[p(F_{\rm obs}; \text{model})] - \sum_{hkl} \ln[p_{\rm Wilson}(F_{\rm obs})]. \qquad (2.2.7.2)$$

LLG-Z score. It is important to note that the LLG depends on the quality of the model and the number of reflections, so the absolute values cannot be compared between different molecular-replacement applications. Instead, the quality of a molecular-replacement solution can be judged by the LLG-Z score, which is defined as the number of standard deviations a score is above the mean score in a particular rotation or translation search.

$$\text{LLG-Z} = \mathrm{LLG} - \langle \mathrm{LLG} \rangle / [(\mathrm{LLG} - \langle \mathrm{LLG} \rangle)^2]^{1/2}. \qquad (2.2.7.3)$$

The translation function Z score (TFZ) for the last component placed in a molecular-replacement search is often a good indicator of the confidence that can be placed in the solution. If TFZ is greater than 8 and there is no translational pseudo-symmetry, the solution is almost always correct.

Detailed descriptions of the background and proper application of molecular-replacement approaches are presented in Chapter 2.3 of *International Tables for Crystallography* Volume B and Chapters 13.2 and 13.3 of the present volume.

2.2.8. Quality indicators for refinement

The last step of a structure determination is the refinement of the model against the observed data. Refinement is in principle a mathematical operation that is applied in order to minimize the discrepancy between the observed structure-factor amplitudes $|F_{\rm obs}|$ and the calculated ones $|F_{\rm calc}|$.

Crystallographic *R* factor, *R*. The crystallographic R factor R is defined as the fractional disagreement between the set of observed structure-factor amplitudes and amplitudes calculated from the structural model. Of course, observed and calculated reflection sets need to be on the same scale.

$$R = \sum_{hkl} \big||F_{\rm obs}| - |F_{\rm calc}|\big| / \sum_{hkl} |F_{\rm obs}|. \qquad (2.2.8.1)$$

Free *R* factor, $R_{\rm free}$. The free R factor $R_{\rm free}$ is defined in the same way as the crystallographic R factor, but it is based on a set of reflections that have been excluded from the refinement (Brünger, 1992). The excluded set of reflections is called the *test set*, while the set of reflections used for refinement is called the *working set*. The test set can be chosen randomly or systematically, either in thin resolution shells or to account for the presence of noncrystallographic symmetry, respectively. In order to minimize the impact on the final model, the test set should be as small as possible. Typically, it contains about 5–10% of the reflections, or at least enough reflections to keep the standard deviation of $R_{\rm free}$ below 1%, but there is no need to use more than 2000 reflections (Kleywegt & Brünger, 1996; Brünger, 1997). The standard deviation of $R_{\rm free}$ has been empirically estimated to be $R_{\rm free}/N^{1/2}$, where N is the number of reflections in the test set (Brünger, 1997). Of course, there may be concerns about the impact of excluding 5–10% of reflections on the final model, but a few final cycles of refinement against the recombined full data set should allay them.

Correlation coefficients CC($F_{\rm obs}$, $F_{\rm calc}$) and CC($I_{\rm obs}$, $I_{\rm calc}$). CC($F_{\rm obs}$, $F_{\rm calc}$) and CC($I_{\rm obs}$, $I_{\rm calc}$) are Pearson linear correlation coefficients [see equation (2.2.2.13)] between observed and model-based calculated structure-factor amplitudes or intensities, respectively, that find use from time to time. One advantage of the use of a correlation coefficient instead of an R factor is that it avoids the problem of scaling the two sets of numbers relative to each other.

2.2.9. Quality indicators for the refined model

In MX, the observable-to-parameter ratio is mostly unfavourable. Therefore, structure refinements are carried out with boundary conditions, constraints and restraints. Constraints reduce the number of parameters which need to be refined, while restraints provide additional information to the refinement procedure that increases the number of observables. A refined model, therefore, has to fulfil not only the criterion that the crystallographic R factor [see equation (2.2.8.1)] is good and that the model fits well to the electron density, but also that it fits well to the restraints used in the refinement procedure.

Real-space residual, RSR. The real-space residual, RSR (Jones *et al.*, 1991), quantifies the discrepancies between the electron-density maps ρ_1, calculated directly from a structural model, and ρ_2, calculated from experimental data. RSR can take the form of a real-space R factor RSRF and of a real-space correlation coefficient RSCC.

$$\mathrm{RSRF} = 2 \sum_{xyz} |\rho_1 - \rho_2| / \sum_{xyz} (\rho_1 + \rho_2). \qquad (2.2.9.1)$$

The sum $\sum_{xyz}$ runs over all grid points of the electron-density maps that are close to the model. A big advantage of RSRF is that is can be calculated on a residue-by-residue basis. It therefore gives a local picture of structure quality. It can also be used throughout model building and refinement in order to follow the improvement of the model locally on a per-residue basis.

RSCC is defined as the Pearson linear correlation coefficient [see equation (2.2.2.13)] between ρ_1 and ρ_2. Everything said about RSRF above applies to RSCC as well.

R.m.s. deviation from ideal of geometric parameter *x*. The root-mean-square deviation of a set of geometric parameters x from their ideal values is defined as

$$\mathrm{r.m.s.d.}(x) = \Big\{ \sum_i [x_i(\text{ideal}) - x_i(\text{observed})]^2 / N \Big\}^{1/2}. \qquad (2.2.9.2)$$

The sum runs over all N instances of the geometric parameter occurring in a structure. The geometric parameters x that are typically considered are bond lengths, bond angles, dihedral angles, chiral volumes, planar groups *etc*. The ideal values for proteins are typically taken from the study of Engh & Huber (1991) and for nucleic acids from Parkinson *et al.* (1996).

***Z* score**. A measure of the likelihood that an individual geometric parameter is correct is given by its Z score. The Z score is defined as the distance of an individual data point of a distribution from the mean of the distribution expressed in standard

Table 2.2.11.1
Definitions of the most commonly used quality indicators

Indicator	Details
Optical resolution, $d_{\rm opt} = [2(\sigma^2_{\rm Patt} + \sigma^2_{\rm sph})]^{1/2}$ (2.2.2.2)	$\sigma_{\rm Patt}$ and $\sigma_{\rm sph}$ are the standard uncertainties of Gaussians fitted to the origin peak of the Patterson function of the diffraction data set and the origin peak of the spherical interference function, respectively
$R_{\rm merge}$ (merging *R* factor), $R_{\rm merge} = \sum_{hkl}\sum_{i} \vert I_i(hkl) - \langle I(hkl)\rangle\vert / \sum_{hkl}\sum_{i} I_i(hkl)$ (2.2.2.4)	$\langle I(hkl)\rangle$ is the mean of the several individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*
$R_{\rm meas}$ or $R_{\rm r.i.m}$ (redundancy-independent merging *R* factor), $R_{\rm r.i.m.} = \sum_{hkl}\{N(hkl)/[N(hkl)-1]\}^{1/2}\sum_{i} \vert I_i(hkl) - \langle I(hkl)\rangle\vert / \sum_{hkl}\sum_{i} I_i(hkl)$ (2.2.2.5)	$\langle I(hkl)\rangle$ is the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*
$R_{\rm p.i.m.}$ (precision-indicating merging *R* factor), $R_{\rm p.i.m.} = \sum_{hkl}\{1/[N(hkl)-1]\}^{1/2}\sum_{i} \vert I_i(hkl) - \langle I(hkl)\rangle\vert / \sum_{hkl}\sum_{i} I_i(hkl)$ (2.2.2.6)	$\langle I(hkl)\rangle$ is the mean of the $N(hkl)$ individual measurements $I_i(hkl)$ of the intensity of reflection *hkl*
$R_{\rm anom}$ (anomalous *R* factor), $R_{\rm anom} = \sum_{hkl} \vert I(hkl) - I(\bar{h}\bar{k}\bar{l})\vert / \sum_{hkl}\langle I(hkl)\rangle$ (2.2.2.12)	$\langle I(hkl)\rangle$ is the mean intensity of the Friedel mates of the reflection *hkl*, or $\frac{1}{2}[I(hkl) + I(\bar{h}\bar{k}\bar{l})]$
$\rm R_{Cullis}$ (Cullis *R* factor for isomorphous-replacement applications), $R_{\rm Cullis} = \sum_{hkl} \big\vert F_{\rm PH} - \vert F_{\rm P} + F_{\rm H}\vert \big\vert / \sum_{hkl} \vert F_{\rm PH} - F_{\rm P}\vert$ (2.2.5.1*b*)	$F_{\rm P}$, $F_{\rm PH}$ and $F_{\rm H}$ are the structure factors for the protein, the heavy-atom derivative and the heavy atoms alone, respectively
$\rm PP_{iso}$ (phasing power for isomorphous-replacement applications), $\mathrm{PP_{iso}} = \sum_{hkl} \vert F_{\rm H}\vert / \sum_{hkl} \big\vert F_{\rm PH} - \vert F_{\rm P} + F_{\rm H}\vert \big\vert$ (2.2.5.3) or, as in the program *SOLVE*, $\mathrm{PP_{iso}} = (\sum_{hkl} \vert F_{\rm H}\vert^2)^{1/2} / (\sum_{hkl} \langle \big\vert F_{\rm PH} - \vert F_{\rm P} + F_{\rm H}\vert \big\vert \rangle^2)^{1/2}$ (2.2.5.4)	$F_{\rm P}$, $F_{\rm PH}$ and $F_{\rm H}$ are the structure factors for the protein, the heavy-atom derivative and the heavy atoms alone, respectively
R (crystallographic *R* factor) and $R_{\rm free}$ (free *R* factor), $R = \sum_{hkl} \big\vert \vert F_{\rm obs}\vert - \vert F_{\rm calc}\vert \big\vert / \sum_{hkl} \vert F_{\rm obs}\vert$ (2.2.8.1)	$R_{\rm free}$ is defined as the crystallographic *R* factor but for a subset of reflections that have been excluded from refinement
RSRF (real-space *R* factor), $\mathrm{RSRF} = 2\sum_{xyz} \vert \rho_1 - \rho_2\vert / \sum_{xyz}(\rho_1 + \rho_2)$ (2.2.9.1)	ρ_1 and ρ_2 are the electron-density maps calculated from the structural model and from the experimental data, respectively
R.m.s.d.'s of geometric parameters *x*, $\mathrm{r.m.s.d}(x) = \{\sum_{i}[x_i(\mathrm{ideal}) - x_i(\mathrm{observed})]^2/N\}^{1/2}$ (2.2.9.2)	x_i are the individual values, ideal or observed, of the geometric parameter *x* and the sum is over all *N* x_i observed. The geometric parameters *x* may be bond lengths, bond angles, dihedral angles, chiral volumes, deviations from planarity *etc*.
DPI (diffraction-component precision index), $\mathrm{DPI} = [N_{\rm atom}/(N_{hkl} - N_{\rm para})]^{1/2} R d_{\rm min} C^{-1/3}$ (2.2.10.1)	$N_{\rm atom}$ is the number of atoms in the structure, N_{hkl} is the number of reflections, $N_{\rm para}$ is the number of refined parameters, *R* is the crystallographic *R* factor, $d_{\rm min}$ is the nominal resolution and *C* is the fractional completeness of the data set

deviations. In the case described here, the mean values of the distribution are the ideal values taken from Engh & Huber (1991) and Parkinson *et al.* (1996).

$$Z(x_i) = [x_i(\text{observed}) - x_i(\text{ideal})]/\sigma(x). \qquad (2.2.9.3)$$

Ideally, the Z score should be 0. A parameter that exhibits a Z score of less than -4 or greater than $+4$ is highly unlikely and calls for attention.

Root-mean-square Z score, r.m.s.-Z. Although r.m.s.d. values [see equation (2.2.9.2)] are still popular for use in judging the quality of refined macromolecular models, a much more useful statistic is the r.m.s. value of a distribution of Z scores or the r.m.s.-Z score.

$$\text{r.m.s.-Z}(x) = \sum_i [Z(x_i)^2/N]^{1/2}. \qquad (2.2.9.4)$$

The sum runs over all N instances of the geometric parameter x occurring in a structure. A very useful property of Z scores is that the r.m.s. values of Z-score distributions should always be 1. Significant deviations from the ideal value indicate potential problems. R.m.s. Z scores are widely used, for instance, in the program *WHAT_CHECK* (Hooft *et al.*, 1996).

R.m.s.d. (NCS). The root-mean-square deviation from crystallographic symmetry between two molecules related by non-crystallographic symmetry (NCS) can be calculated from a superposition of the two molecules. It is defined as

$$\text{r.m.s.d.(NCS)} = \left(\sum_i d_i^2/N\right)^{1/2}. \qquad (2.2.9.5)$$

The sum runs over N equivalent atom pairs with d_i being the distance between the two equivalent atoms after superposition.

2.2.10. Error estimation for the refined model

An important quality indicator for a refined model is the coordinate uncertainty. Short of full-matrix inversion, which is the standard procedure in small-molecule crystallography but which is applicable only in exceptional cases for macromolecules, some methods have been devised for estimating of the overall coordinate uncertainty.

Error estimation according to Luzzati. For most macromolecular structure determinations, atomic standard uncertainties are not available. However, Luzzati (1952) devised a method of estimating the average positional error of a structure. Under the assumption that the atomic positional errors follow a normal distribution, the average error can be estimated by comparing a plot of the crystallographic R factor [see equation (2.2.8.1)] *versus* the reciprocal resolution (or 2 sin θ/λ) with pre-computed theoretical curves for different average errors. A more recent – and probably better – approach is to use the free R factor instead of the crystallographic R factor.

SigmaA- (σ_A)-type error estimation. A slightly better estimate of the average positional error of a structure can be obtained by plotting the natural logarithm of the parameter σ_A *versus* $(\sin\theta/\lambda)^2$ (Read, 1986). The slope of a straight line fitted to the plot provides an estimate of the average positional error of the structure. The parameter σ_A assumes normally distributed positional errors and takes model incompleteness into account as well.

Diffraction-component precision index, DPI. The diffraction-component precision index DPI is an empirical parameter describing the overall coordinate uncertainty of a structure (Cruickshank, 1999*a*,*b*). For an atom with an isotropic displacement parameter of average value (B_{avg}), it is defined as

$$\text{DPI} = [N_{\text{atom}}/(N_{hkl} - N_{\text{para}})]^{1/2} R d_{\min} C^{-1/3}, \qquad (2.2.10.1)$$

where N_{atom} is the number of atoms included in the refinement, N_{hkl} is the number of reflections included in the refinement, N_{para} is the number of refined parameters, R is the crystallographic R factor, $d_{\min}$ is the nominal resolution of the data included in the refinement and C is the data completeness. The free R factor R_{free} is sometimes used instead of the crystallographic R factor R to calculate the DPI. In this case $(N_{hkl} - N_{\text{para}})$ is replaced by N_{free}, which is the number of reflections used for R_{free} calculation.

The Cruickshank formula for DPI has been recast into several other forms, including

$$\text{DPI} = \sigma(x, B_{\text{avg}}) = 0.18(1 + sc)^{1/2} V_M^{-1/2} R_{\text{free}} d_{\min}^{5/2} C^{-5/6} \qquad (2.2.10.2)$$

(Blow, 2002), where sc is the solvent fraction ($= N_{\text{solv}}/N_{\text{atom}}$, where N_{solv} is the number of atoms that are solvent) and V_M is the Matthews parameter (Matthews, 1968). The utility of this latter formula in guiding the design of the data-collection experiment to achieve a specified target coordinate uncertainty has been demonstrated (Fisher *et al.*, 2008).

2.2.11. The most commonly used quality indicators

A summary of the most commonly used quality indicators and their definitions is presented in Table 2.2.11.1 for ready reference.

We gratefully acknowledge the contributions of Kevin Cowtan (York, UK), Kay Diederichs (Konstanz, Germany), Phil Evans (Cambridge, UK), John Helliwell (Manchester, UK), Randy Read (Cambridge, UK), George Sheldrick (Göttingen, Germany) and Tom Terwilliger (Los Alamos, USA).

References

Arnberg, L., Hovmöller, S. & Westman, S. (1979). *On the significance of 'non-significant' reflexions*. *Acta Cryst.* A**35**, 497–499.

Blow, D. M. (2002). *Rearrangement of Cruickshank's formulae for the diffraction-component precision index*. *Acta Cryst.* D**58**, 792–797.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method*. *Acta Cryst.* **12**, 794–802.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1997). *Free R value: cross-validation in crystallography*. *Methods Enzymol.* **277**, 366–396.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography*. *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1999). *Error estimation and bias correction in phase-improvement calculations*. *Acta Cryst.* D**55**, 1555–1567.

Cowtan, K. & Main, P. (1998). *Miscellaneous algorithms for density modification*. *Acta Cryst.* D**54**, 487–493.

Cowtan, K. D. & Main, P. (1996). *Phase combination and cross validation in iterated density-modification calculations*. *Acta Cryst.* D**52**, 43–48.

Cruickshank, D. W. J. (1999*a*). *Remarks about protein structure precision*. *Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (1999*b*). *Remarks about protein structure precision. Erratum*. *Acta Cryst.* D**55**, 1108.

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1961). *The structure of haemoglobin. VIII. A three-dimensional Fourier synthesis at 5.5 Å resolution: determination of the phase angles*. *Proc. R. Soc. London Ser. A*, **265**, 15–38.

Diederichs, K. (2006). *Some aspects of quantitative analysis and correction of radiation damage*. *Acta Cryst.* D**62**, 96–101.

Diederichs, K. (2010). *Quantifying instrument errors in macromolecular X-ray data sets*. *Acta Cryst.* D**66**, 733–740.

Diederichs, K. & Karplus, P. A. (1997*a*). *Improved R-factors for diffraction data analysis in macromolecular crystallography*. *Nat. Struct. Biol.* **4**, 269–275.

Diederichs, K. & Karplus, P. A. (1997*b*). *Improved R-factors for diffraction data analysis in macromolecular crystallography. Erratum*. *Nat. Struct. Biol.* **4**, 592.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement*. *Acta Cryst.* A**47**, 392–400.

Evans, P. (2006). *Scaling and assessment of data quality*. *Acta Cryst.* D**62**, 72–82.

Fisher, S. J., Helliwell, J. R., Khurshid, S., Govada, L., Redwood, C., Squire, J. M. & Chayen, N. E. (2008). *An investigation into the protonation states of the C1 domain of cardiac myosin-binding protein C*. *Acta Cryst.* D**64**, 658–664.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Substructure search procedures for macromolecular structures*. *Acta Cryst.* D**59**, 1966–1973.

Hauptman, H. & Karle, J. (1953). *Solution of the phase problem. I. The centrosymmetric crystal*. Am. Crystallogr. Assoc. Monograph No. 3. Dayton, Ohio: Polycrystal Book Service.

Hirshfeld, F. L. & Rabinovich, D. (1973). *Treating weak reflexions in least-squares calculations*. *Acta Cryst.* A**29**, 510–513.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures*. *Nature (London)*, **381**, 272.

Howell, P. L. & Smith, G. D. (1992). *Identification of heavy-atom derivatives by normal probability methods*. *J. Appl. Cryst.* **25**, 81–86.

James, R. W. (1948). *False detail in three-dimensional Fourier representations of crystal structures*. *Acta Cryst.* **1**, 132–134.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models*. *Acta Cryst.* A**47**, 110–119.

Kabsch, W. (1988). *Evaluation of single-crystal X-ray diffraction data from a position-sensitve detector*. *J. Appl. Cryst.* **21**, 916–924.

Kabsch, W. (1993). *Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants*. *J. Appl. Cryst.* **26**, 795–800.

Kabsch, W. (2010). *XDS*. *Acta Cryst.* D**66**, 125–132.

Kleywegt, G. & Brünger, A. T. (1996). *Checking your imagination: applications of the free R value*. *Structure*, **4**, 897–904.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *BALBES: a molecular-replacement pipeline*. *Acta Cryst.* D**64**, 125–132.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines*. *Acta Cryst.* **5**, 802–810.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software*. *J. Appl. Cryst.* **40**, 658–674.

Matthews, B. W. (1968). *Solvent content of protein crystals*. *J. Mol. Biol.* **33**, 491–497.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode*. *Methods Enzymol.* **276**, 307–326.

Panjikar, S. & Tucker, P. A. (2002). *Phasing possibilities using different wavelengths with a xenon derivative*. *J. Appl. Cryst.* **35**, 261–266.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures*. *Acta Cryst.* D**52**, 57–64.

Pflugrath, J. W. (1999). *The finer things in X-ray diffraction data collection*. *Acta Cryst.* D**55**, 1718–1725.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors*. *Acta Cryst.* A**42**, 140–149.

Read, R. J. (2001). *Pushing the boundaries of molecular replacement with maximum likelihood*. *Acta Cryst.* D**57**, 1373–1382.

Rodgers, J. L. & Nicewander, W. A. (1988). *Thirteen ways to look at the correlation coefficient*. *Am. Stat.* **42**, 59–66.

Rodriguez, D. D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Uson, I. (2009). *Crystallographic ab initio protein structure solution below atomic resolution*. *Nat. Methods*, **6**, 651–653.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit*. *Acta Cryst.* **15**, 24–31.

Schneider, T. R. & Sheldrick, G. M. (2002). *Substructure solution with SHELXD*. *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (2002). *Macromolecular phasing with SHELXE*. *Z. Kristallogr.* **217**, 644–650.

Sheldrick, G. M. (2008). *A short history of SHELX*. *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2010). *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*. *Acta Cryst.* D**66**, 479–485.

Stenkamp, R. E. & Jensen, L. H. (1984). *Resolution revisited: limit of detail in electron density maps*. *Acta Cryst.* A**40**, 251–254.

Terwilliger, T. C. (2001). *Map-likelihood phasing*. *Acta Cryst.* D**57**, 1763–1775.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard*. *Acta Cryst.* D**65**, 582–601.

Terwilliger, T. C. & Berendzen, J. (1999). *Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods*. *Acta Cryst.* D**55**, 501–505.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model*. *Acta Cryst.* D**55**, 191–205.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification*. *J. Appl. Cryst.* **28**, 347–351.

Weeks, C. M., De Titta, G. T., Miller, R. & Hauptman, H. A. (1993). *Application of the minimal principle to peptide structures*. *Acta Cryst.* D**49**, 179–181.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications*. *Acta Cryst.* A**50**, 210–220.

Weiss, M. S. (2001). *Global indicators of X-ray data quality*. *J. Appl. Cryst.* **34**, 130–135.

Weiss, M. S. & Hilgenfeld, R. (1997). *On the use of the merging R factor as a quality indicator for X-ray data*. *J. Appl. Cryst.* **30**, 203–205.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities*. *Acta Cryst.* **2**, 318–321.

International Tables for Crystallography (2012). Vol. F, Chapter 3.1, pp. 75–91.

PART 3. TECHNIQUES OF MOLECULAR BIOLOGY

Chapter 3.1. Preparing recombinant proteins for X-ray crystallography

S. H. Hughes and A. M. Stock

3.1.1. Introduction

Preparing protein crystals appropriate for X-ray diffraction usually requires a considerable amount of highly purified protein. When crystallographic methods were first developed, the practitioners of the art were compelled to study proteins that could be easily obtained in large quantities in relatively pure form; the first proteins whose structures were solved by crystallographic methods were myoglobin and haemoglobin. Unfortunately, some of the most interesting proteins are normally present in relatively small amounts, which, while it did not prevent crystallographers from dreaming about their structures, prevented any serious attempts at crystallization. Recombinant DNA techniques changed the rules: it is now possible to instruct a variety of cells and organisms to make large amounts of almost any protein chosen by the investigator. Not only can specific proteins be expressed in large quantities, recombinant proteins can be modified in ways that make the task of the crystallographer simpler and can, in some cases, dramatically improve the quality of the resulting crystals. It is not our intention in writing this chapter to provide either a methods manual for those interested in expressing a particular protein or a complete compendium of the available literature. The literature is vast and complex, and, as we will discuss, the problems associated with expressing a particular protein are often idiosyncratic, making it difficult to provide a simple, comprehensive, methodological guide. What we intend is to discuss issues (and problems) relevant to choosing methods appropriate for preparing recombinant proteins for X-ray crystallography. In this way, we hope to help readers understand both the extant problems and the available solutions, so that, armed with a general understanding of the issues, they can more easily confront a variety of specific projects.

Fortunately, there are a large number of additional resources available to those who are interested in expressing and purifying recombinant proteins, but lack the expertise. These include numerous methods books (*e.g.* on molecular biology: Sambrook *et al.*, 1989; Ausubel *et al.*, 1995; on protein purification: Abelson & Simon, 1990; Scopes, 1994; Bollag *et al.*, 1996), useful reviews of the literature (cited throughout), formal courses (such as those offered by Cold Spring Harbor Laboratory), meetings (*i.e.* IBC's International Conference on Expression Technologies, Washington DC, 1997) and a specialized journal (*Protein Expression and Purification*). The pace of methodological development is rapid, and company catalogues, publications and web pages can provide extensive, useful, up-to-date information. In many cases, a convenient source of information is a nearby researcher whose own research depends on expressing and purifying recombinant proteins. Those who are serious about preparing recombinant proteins for crystallography, but have little or no experience, are strongly urged to avail themselves of these resources. In many cases the help of a knowledgeable colleague is the most valuable resource. In general, the literature provides a much better guide to what will work than what will fail; quite often, in designing a good strategy to produce a recombinant protein that is suitable for crystallography, it is more important to understand the potential pitfalls. Discussion with an experienced colleague is usually the best way to avoid the most obvious errors.

Section 3.1.2 gives an overview of the problem, Section 3.1.3 discusses engineering an expression construct, Section 3.1.4 discusses expression systems, Section 3.1.5 discusses protein purification and Section 3.1.6 discusses the characterization of the purified product.

3.1.2. Overview

The idea that underlies the problem of expressing large amounts of a recombinant protein is straightforward: prepare a DNA segment that, when introduced into an appropriate host, will cause the abundant expression of the relevant protein. However, as the saying goes, 'The devil is in the details.' Not only is it necessary to design the appropriate DNA segment, but also to introduce it into an appropriate host such that the host retains and faithfully replicates the DNA. The DNA segment must contain all of the elements necessary for high-level RNA expression; moreover, the RNA, when expressed, must be recognized by the translational machinery of the host. The recombinant protein, once expressed, needs to be properly folded either by the host or, if not properly folded in the host, by the experimentalist. If the protein is subject to post-translational modifications (cleavage, glycosylation, phosphorylation *etc.*) and the experimentalist wishes to retain these modifications, the appropriate signals must be present and the chosen host must also be capable of recognizing the signals. Once the recombinant protein is expressed, assuming it is reasonably stable in the chosen host, the protein must be purified; as we will discuss, recombinant proteins can be modified to simplify purification. Once purified, the quality of the protein preparation must be evaluated to ensure it is both relatively homogeneous and monodisperse.

While this chapter will be limited to discussions of the basic strategies for creating an expression vector, expressing the protein and purifying and characterizing the product, molecular biological methods can be used in other ways that are relevant to crystallography. In some cases, a protein in its natural form is not suitable for crystallization. Crystallographers have long used proteolytic digestion and/or glycolytic digestion to produce proteins suitable for crystallization from ones that are not. Such techniques have been used to good effect on recombinant proteins; however, the ability to modify the segment encoding the protein makes it possible to alter the protein in a variety of ways beyond simple enzymatic digestions. Specific examples of such applications are described in Chapter 4.3.

Unfortunately, no single strategy for producing proteins for crystallization appears to be universally successful. Any parti-

Copyright © 2012 International Union of Crystallography

cular protocol has the potential for displaying undesirable behaviour at any step during the process of expression, purification or crystallization. It is important to distinguish major and minor problems. If the problems are serious, it is often better to try an alternative strategy than to struggle with an inappropriate system. Because it is usually difficult to predict what will work and what will not, often the most expedient route to successful expression of a protein for crystallization is the simultaneous pursuit of several expression strategies with multiple protein expression constructs.

3.1.3. Engineering an expression construct

3.1.3.1. Choosing an expression system

The first step in developing an expression strategy is the choice of an appropriate expression system, and this decision is critical. As we will discuss briefly below, the rules and/or sequences necessary to express RNA and proteins in *E. coli*, yeast and insect cells (baculoviruses) differ to a greater or lesser extent from those used in higher eukaryotes, and there are considerable differences in the post-translational modifications of proteins in these different systems or organisms. Quite often the protein chosen for investigation comes from a higher eukaryote or from a virus that replicates in higher eukaryotes. The experimentalist prefers to obtain large amounts of the protein (>5–10 mg) to set up crystallization trials. In theory, one simple solution is to use a closely related host to express the protein of interest. While it is possible to produce large amounts of proteins in cultured animal cells (and in some cases in transgenic animals), the difficulties and expense of these approaches usually prevent their use for most projects that require large amounts of highly purified recombinant protein.

In general, prokaryotic (*E. coli*) expression systems are the easiest to use in terms of the preparation of the expression construct, the growth of the recombinant organism and the purification of the resulting protein. Additionally, they allow for relatively easy incorporation of selenomethionine into the recombinant protein (Hendrickson *et al.*, 1990), which is an important consideration for crystallographers intending to use multiple anomalous dispersion (MAD) phasing techniques. However, the differences between *E. coli* and higher eukaryotes means that, in some cases, the recombinant protein must be modified to permit successful expression in *E. coli*, and the available *E. coli* expression systems cannot produce many of the post-translational modifications made in higher eukaryotes. As one moves along the evolutionary path from *E. coli* to yeast, to baculovirus and finally to cultured mammalian cells, the problems associated with producing the protein in its native state are simpler, while the problems associated with expressing large amounts of material quickly, simply and cheaply in an easy-to-purify form become more difficult. In Section 3.1.4, we will consider each of these expression systems in turn; first we will briefly discuss, in a general way, how the relevant genes or cDNA strands are obtained and how an expression system is designed.

3.1.3.2. Creating an expression construct

The first step in preparing an expression system is obtaining the gene of interest. This is not nearly as daunting a task as it once was; an intense effort is now being directed at genome sequencing and the preparation of cDNA clones from a number of prokaryotic and eukaryotic organisms. There are also a large number of cloned viral genes and genomes. This means that, in most cases, an appropriate gene or cDNA can be obtained without the need to prepare a clone *de novo*. If the nucleic sequence is available, but the corresponding cloned DNA is not, it is usually a simple matter to prepare the desired DNA clone using the polymerase chain reaction (PCR). If the relevant genomic or cDNA clone is not available and there is no obvious way to obtain it, there are established techniques for obtaining the desired clone; however, these methods are often tedious and labour intensive. They also constitute a substantial field in their own right and, as such, lie beyond the scope of this chapter (for an overview, see Sambrook *et al.*, 1989).

In higher eukaryotes, most mRNA strands are spliced. With minor exceptions, mRNA strands are not spliced in *E. coli*. In yeast, the splicing rules do not match those used in higher eukaryotes. If one expects to express a protein from a higher eukaryote in one of these systems, a cDNA must be prepared or obtained. Because some introns are large, cDNA clones are often used as the basis of expression constructs in baculovirus systems, as well as in cultured insect and mammalian cells.

In all subsequent discussions, we will assume that the experimentalist possesses both a cDNA that encodes the protein that will be expressed and an accurate sequence. If a genomic clone is available, it can be converted to cDNA form by PCR methods or by using a retroviral vector. Retroviral vectors, by nature of their life cycle, will take a gene through an RNA intermediate, thus removing unwanted introns (Shimotohno & Temin, 1982; Sorge & Hughes, 1982). If a good sequence is not available, one should be prepared. In general, expression constructs are based, more or less exclusively, on the coding region of the cDNA. The flanking 5′ and 3′ untranslated regions are not usually helpful, and if these untranslated regions are included in an expression construct, they can, in some cases, interfere with transcription, translation or both. With some knowledge of the organization of the protein, it is sometimes helpful to express portions of a complex protein for crystallization. This will be discussed in more detail later in this chapter and in Chapter 4.3.

Optimizing the expression of the protein is extremely important. The amount of effort required to get an expression system to produce twice as much protein is usually less than that required to grow twice as much of the host; moreover, the effort to purify a recombinant protein is inversely related to its abundance, relative to the proteins of the host. There are specific rules for expressing a recombinant protein in the different host–vector systems; these will be discussed in the context of using various hosts (*E. coli*, yeast, baculoviruses and cultured insect and mammalian cells).

Although the precise nature of the modifications necessary to obtain efficient expression of a protein is host dependent, the tools used to produce the modified cDNA and link it to an appropriate expression plasmid or other vector are reasonably standard. In recent years, PCR has become the method of choice for manipulation of DNA; it is a relatively easy and rapid method for altering DNA segments in a variety of useful ways (Innis *et al.*, 1990; McPherson *et al.*, 1995). For most construction projects, the ends of the cDNA are modified, using PCR with appropriate oligonucleotide primers that have been designed to introduce useful restriction sites and/or elements essential for efficient transcription and/or translation. Since it can often be advantageous to try the expression of a given protein construct in a number of different vectors, it is useful to incorporate carefully chosen restriction sites that will enable the fragment to be inserted simultaneously, or transferred seamlessly, into different plasmids or other vectors (Fig. 3.1.3.1). PCR can also be used to create mutations in the interior of the cDNA. For some projects

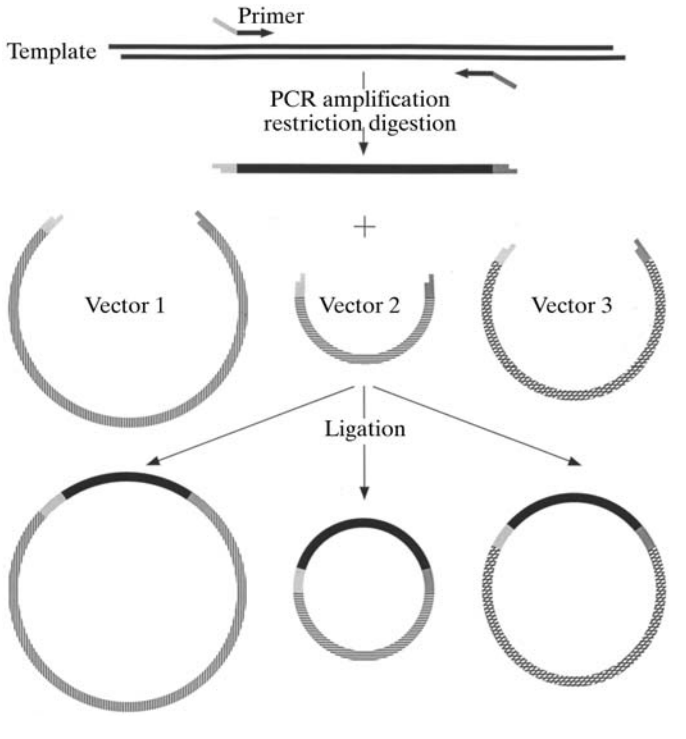


Figure 3.1.3.1
Creating an expression construct. PCR can be used to amplify the coding region of interest, providing that a suitable template is available. PCR primers should be designed to contain one or more restriction sites that can be conveniently used to subclone the fragment into the desired expression vector. It is often possible to choose vectors and primers such that a single PCR product can be ligated to multiple vectors. The ability to test several expression systems simultaneously is advantageous, since it is impossible to predict which vector/host system will give the most successful expression of a specific protein.

where large-scale mutagenesis is planned, other mutagenic techniques are particularly helpful (for example, site-directed cassette mutagenesis using *Bsp* MI or a related enzyme; Boyer & Hughes, 1996). Ordinarily, however, these alternative strategies are only useful if a relatively large number of mutants are needed for the project.

If PCR is used either to modify the ends of a DNA segment or to introduce specific mutations within a segment, it should be remembered that the PCR can introduce unwanted mutations. PCR conditions should be chosen to minimize the risk of introducing unwanted mutations (start with a relatively large amount of template DNA, limit the number of amplification cycles, use relatively stringent conditions for hybridization of the primers, choose solution conditions that reduce the number of errors made in copying the DNA and use enzymes with good fidelity, such as *Pfu* or others that have proofreading capabilities). It is also important to sequence all of the DNA pieces generated by PCR after they have been cloned.

3.1.3.3. Addition of tags or domains

In some cases it is useful to add a small peptide tag or a larger protein to either the amino or carboxyl terminus of the protein of interest (Nilsson *et al.*, 1992; LaVallie & McCoy, 1995). As will be discussed in more detail below, such fused elements can be used for affinity chromatography and can greatly simplify the purification of the recombinant protein. In addition to aiding purification, some protein domains used as tags, such as the maltose-binding protein, thioridazine, and protein A, can also act as molecular chaperones to aid in the proper folding of the recombinant protein (LaVallie *et al.*, 1993; Samuelsson *et al.*, 1994; Wilkinson *et al.*, 1995; Richarme & Caldas, 1997; Sachdev & Chirgwin, 1998). Tags range in size from several amino acids to tens of kilodaltons. Numerous tags [including hexahistidine (His$_6$), biotinylation peptides and streptavidin-binding peptides (Strep-tag), calmodulin-binding peptide (CBP), cellulose-binding domain (CBD), chitin-binding domain (CBD), glutathione S-transferase (GST), maltose-binding protein (MBP), protein A domains, ribonuclease A S-peptide (S-tag) and thioridazine (Trx)] have already been engineered into expression vectors that are commercially available. Additional systems are constantly being introduced. While these systems provide some advantages, there are also drawbacks, including expense, which can be considerable when both affinity purification and specific proteolytic removal of the tag are performed on a large scale.

If a sequence tag or a fusion protein is added to the protein of interest, one problem is solved but another is created, *i.e.* whether or not to try to remove the fused element. During the past year, there have been numerous reports of crystallization of proteins containing His-tags, but there are also unpublished anecdotes about cases where removal of the tag was necessary to obtain crystals. In a small number of cases, additional protein domains present in fusion proteins appear to have aided crystallization (see Chapter 4.3). Experiences with tags appear to be protein specific. There are a number of relevant issues, including the protein, the tag and the length and composition of the linker that joins the two. If the tag is to be removed, it is usually necessary to use a protease. To avoid unwanted cleavage of the desired protein, 'specific' proteases are usually used. When the expression system is designed, the tag or fused protein is separated from the desired protein by the recognition site for the protease. While this procedure sounds simple and straightforward, and has, in some cases, worked exactly as outlined here, there are a number of potential pitfalls. Proteases do not always behave exactly as advertised, and there can be unwanted cleavages in the desired product. Since protease cleavage efficiency can be quite sensitive to structure, it may be more difficult to cleave the fusion joint than might be expected. Unless cleavage is performed with an immobilized protease, additional purification is necessary to separate the protease from the desired protein product. A variation of the classic tag-removal procedure is provided by a system in which a fusion domain is linked to the protein of interest by a protein self-cleaving element called an intein (Chong *et al.*, 1996, 1997).

3.1.4. Expression systems

3.1.4.1. E. coli

If the desired protein does not have extensive post-translational modifications, it is usually appropriate to begin with an *E. coli* host–vector system (for an extensive review of expression in *E. coli*, see Makrides, 1996). Both plasmid-based and viral-based (M13, λ *etc.*) expression systems are available for *E. coli*. Although viral-based vector systems are quite useful for some purposes (expression cloning of cDNA strands, for example), in general, for expression of relatively large amounts of recombinant protein, they are not as convenient as plasmid-based expression systems. Although there are minor differences in the use of viral expression systems and plasmid-based systems, the rules that govern the design of the modified segment are the same and we will discuss only plasmid-based systems. We will first consider general issues related to design of the plasmid, then continue with a discussion of fermentation conditions, and finally address some of the problems commonly encountered and potential solutions.

Basically, a plasmid is a small circular piece of DNA. To be retained by *E. coli*, it must contain signals that allow it to be successfully replicated by the host. Most of the commonly used *E. coli* expression plasmids are present in the cell in multiple copies. Simply stated, in the selection of *E. coli* containing the plasmid, the plasmids carry selectable markers, which usually confer resistance to an antibiotic, typically ampicillin and/or kanamycin. Ampicillin resistance is conferred by the expression of a β-lactamase that is secreted from cells and breaks down the antibiotic. It has been found that, in typical liquid cultures, most of the ampicillin is degraded by the time cells reach turbidity (approximately 10^7 cells ml^{-1}), and cells not harbouring plasmids can overgrow the culture (Studier & Moffatt, 1986). For this reason, kanamycin resistance is being used as the selectable marker in many recently constructed expression plasmids.

There are literally dozens, if not hundreds, of expression plasmids available for *E. coli*, so a comprehensive discussion of the available plasmids is neither practical nor useful. Fortunately, this broad array of choices means that considerable effort has been expended in developing *E. coli* expression systems that are efficient and easy to use (for a concise review, see Unger, 1997). In most cases, it is possible to find expression and/or fermentation conditions that result in the production of a recombinant protein that is at least several per cent of the total *E. coli* protein. This should result in the expression of greater than 5 mg of recombinant protein per litre of culture, making the scale of fermentation reasonable and the job of purification relatively simple.

Broadly speaking, *E. coli* expression systems are either constitutive (that is, they always express the encoded protein) or inducible, in which case a specific change in the culture conditions is necessary to induce the expression of the recombinant protein. As is often the case, both systems have advantages and disadvantages, and both systems have been successfully used to generate recombinant protein for X-ray crystallographic experiments. There is no question that constitutive systems are simple and convenient. However, the high-level expression of even a relatively benign recombinant protein usually puts the *E. coli* host at a selective disadvantage. Unless precautions are taken, the growth and repeated passage of *E. coli* carrying a constitutive expression plasmid tend to select for variants that express lower (and sometimes much lower) levels of the desired recombinant protein than were seen when the clone was first prepared. This can be avoided by storing the stock as plasmid DNA and regularly preparing fresh transformants.

If the desired protein is toxic to *E. coli* (as are a substantial number of recombinant proteins), then an inducible system is required. There are several considerations when choosing an inducible system. The method used to induce the expression of the protein should be compatible with the scale required to produce the recombinant protein. For example, inducible systems which use the bacteriophage λ $p_{\rm L}$ promoter and the temperature-sensitive repressor CI857ts require a temperature shift from approximately 30 to 42 °C. This can be done quite conveniently in small cultures, but it is much more difficult to achieve a rapid shift of temperature if *E. coli* are grown in batches larger than 10 l. Inducible expression systems based on the *lac* repressor are usually induced with isopropyl-β-D-thiogalactopyranoside (IPTG). The cost of this gratuitous inducer is not an issue when *E. coli* are grown in small cultures; however, in large-scale fermentations, the costs of the inducer are nontrivial. Despite this caveat, expression systems controlled by the *lac* repressor are commonly used. In the original *lac*-based inducible expression systems, the *lac* operator/promoter was located on the plasmid, proximal to the 5′ end of the insert. Because expression plasmids are present in multiple copies in *E. coli*, the *lac* repressor must be overexpressed to a substantial degree for it to be present in sufficient quantity to control a plasmid-borne operon. Even if a highly expressed *lac* repressor gene (*lacI*$^{\rm q}$, which produces approximately ten times as much repressor than does the wild type) is expressed from a single chromosomal copy (*i.e.*, provided by the host strain rather than by the vector), repression is rarely complete, and some constitutive expression is generally observed, with only moderately increased levels of expression achieved upon induction. Note that the same plasmid constructs will often give different levels of expression of the plasmid-borne gene in different host strains because of the nature of the *lac* repressor gene (wild-type or *lacI*$^{\rm q}$).

Better control of induction can usually be obtained using a T7 polymerase expression system in a specifically designed vector–host strain pair (Tabor & Richardson, 1985; Studier & Moffatt, 1986; Studier *et al.*, 1990). In such systems, a *lac*-controlled operon that encodes the bacteriophage T7 RNA polymerase is embedded in the genome of the *E. coli* host and is, as a consequence, present in the cell in only one copy. Induction with IPTG leads to the synthesis of the T7 RNA polymerase, which recognizes a promoter sequence that is different from the sequence recognized by *E. coli* RNA polymerase. If the *E. coli* host that carries the T7 RNA polymerase under the control of *lac* also carries a multicopy plasmid, in which the gene of interest is linked to a T7 promoter, the T7 RNA polymerase efficiently produces mRNA from the plasmid; this usually leads to the production of a large amount of the desired recombinant protein. *E. coli* strains that carry a *lac*-inducible T7 RNA polymerase are readily available, as are the corresponding expression plasmids that carry T7 promoters. Some such *E. coli* strains have been specifically engineered so that the expression of the T7 RNA polymerase (and, by extension, the expression of the gene of interest on the plasmid) is tightly regulated (Studier *et al.*, 1990); these strains are particularly useful for expressing recombinant proteins that are toxic to the *E. coli* host. A recent variation on this system uses an *E. coli* strain in which the T7 RNA polymerase gene is under control of the NaCl-induced *proU* promoter (Bhandari & Gowrishankar, 1997). The same plasmids used for other T7 systems can be used with this *E. coli* strain. The osmo-regulated system has the advantages of requiring a much less expensive inducer and, in at least some cases where inclusion-body formation is a problem, of producing higher levels of soluble protein.

Adaptations of a cDNA may be necessary for high-level expression in *E. coli* (Table 3.1.4.1). Although the genetic code is universal, the signals necessary for transcription of RNA and translation of proteins are not. Most *E. coli* expression plasmids contain the recognition/regulation sites necessary for controlling RNA transcription; the signals necessary to initiate translation are not always included in expression plasmids. In *E. coli*, the initiation of translation requires not only an appropriate initiation codon (usually AUG, occasionally GUG), but also a special element, the Shine–Dalgarno sequence, just 5′ of the initiator AUG (Gold *et al.*, 1981; Ringquist *et al.*, 1992). In *E. coli*, the first step in translation involves the binding of the 30S ribosomal subunit and the initiator fMet-tRNA to the mRNA. The Shine–Dalgarno sequence is complementary to the 3′ end of the 16S

Table 3.1.4.1
Strategies for improving expression in *E. coli*

See text for details.

Factor limiting expression	Possible solution
Transcription and/or translation initiation sites	Use vectors with optimized promoter regions
Toxicity	Use inducible expression systems Use mutagenesis to eliminate enzymatic activity Express a domain of the protein
Rare codons	Use plasmids that co-express corresponding tRNA strands Use mutagenesis to optimize codons
Proteolysis	Use protease-deficient host strains Use N-end rules to avoid degradation
Inclusion-body formation	Express the protein as a fusion Co-express chaperone proteins Grow cells at lower temperatures

RNA found in the 30S subunit. Eukaryotic mRNAs do not contain Shine–Dalgarno sequences. Some *E. coli* expression plasmids carry a Shine–Dalgarno sequence, others do not. If one is not present in the plasmid, it must be introduced when the cDNA sequence is modified for introduction into the expression plasmid.

The Shine–Dalgarno sequence needs to be positioned in close proximity to the ATG. Ideally, the nucleotide that pairs with C1535 of the 16S RNA should be positioned eight nucleotides upstream of the A of the initiation codon, although a range of 4–14 nucleotides is tolerated (Gold *et al.*, 1981). If the Shine–Dalgarno sequence is supplied by the plasmid, the restriction enzyme recognition site used to join the cDNA to the plasmid must be quite close to the ATG. Expression systems have been developed in which the restriction site used for creating the expression system includes the initiator ATG. Many expression plasmids are available that have *Nde*I (CAT$\underline{\text{ATG}}$) or *Nco*I (CC$\underline{\text{ATG}}$G) recognition sites at the initiator ATG, which makes it possible to move only the coding region of the cDNA into the expression plasmid. Note, however, that if the *Nco*I site is used, retaining the *Nco*I site in the final construction specifies the first base of the second codon. This limits the choices for the second amino acid in the recombinant protein. *Nde*I does not have this limitation.

The termini of proteins influence their susceptibility to degradation by cellular proteases, most notably ClpA. The N-end rule for bacteria is that proteins in which the N-terminal amino acid is Phe, Leu, Trp, Tyr, Arg or Lys are unusually susceptible to proteolysis (Tobias *et al.*, 1991). (The stability of proteins beginning with Pro has not been determined.) These amino acids (as well as others) seem to impart instability in eukaryotic cells as well. Thus in most cases, if one is expressing intact proteins, the N-terminal amino acid of the native sequence will not generally present a problem. Furthermore, under most circumstances, generation of proteins with N-end-rule amino acids at their N-termini are unlikely, since all proteins are initiated with methionine, and while N-terminal methionines are sometimes removed, the specificity of *E. coli* aminopeptidase is such that methionines adjacent to N-end-rule amino acids are not removed with high efficiency (Hirel *et al.*, 1989). Note that methionine removal sometimes occurs, though to a lesser degree, if the penultimate residue is Asn, Asp, Leu or Ile. Thus, Leu residues should probably be avoided at the second position. It is also possible to generate termini containing N-end-rule amino acids by endoproteolytic cleavage; thus it might be advantageous to avoid these amino acids at the beginnings of proteins where unstructured ends are suspected.

Codon usage can influence expression levels. Although it is not something that should routinely be considered in the initial stages of a project, it is a factor that should be kept in mind if no or low levels of expression are observed. Although the genetic code is universal, it is also degenerate: twenty amino acids are specified by 61 codons. Most amino acids are specified by more than one codon; in many cases, some of the codons are used more often (and translated more efficiently) than others. Unfortunately, there are substantial differences in codon preference/usage in prokaryotes and eukaryotes (Zhang *et al.*, 1991). In *E. coli*, codon usage reflects the abundance of the cognate tRNA strands, and poorly expressed genes tend to contain a higher frequency of rare codons (De Boer & Kastelein, 1986). Although a number of theories have been proposed, prediction of the adverse effects of rare codons on the expression of any given sequence is not currently feasible. Factors such as the position of the codons, their clustering or dispersity and the RNA secondary structure may all contribute to levels of expression (Goldman *et al.*, 1995; Kane, 1995). In many instances, *E. coli* do make relatively large amounts of recombinant protein from mRNA strands that contain a number of rare codons (Ernst & Kawashima, 1988; Lee *et al.*, 1992). But in other cases, optimizing codon usage (Hernan *et al.*, 1992; Mohsen & Vockley, 1995) or co-expressing low abundance tRNAs (Brinkmann *et al.*, 1989; Del Tito *et al.*, 1995) has improved the level of expression of recombinant proteins. Since oligonucleotides 50–75 bases long can be synthesized relatively easily, it is possible to create relatively large synthetic cDNA strands or genes that have optimal codon usage. An alternative strategy is to take advantage of plasmids that have been constructed for co-expression of low abundance tRNA strands [$\text{tRNA}^{\text{Arg(AGA/AGG)}}$ and $\text{tRNA}^{\text{Ile(AUA)}}$] (Schenk *et al.*, 1995; Kim *et al.*, 1998). Fortunately, these strategies are not usually necessary; before attempting to optimize codon usage, one should first ask whether the natural sequence can be expressed efficiently.

Once plasmid constructs have been created and strains have been assembled, it is important that they be properly stored. Although it is possible to persuade *E. coli* to make large amounts of recombinant protein, it should be remembered that this is an artificial situation chosen by the investigator, not the *E. coli* host. As such, it behoves the experimentalist to pay careful attention to the host; *E. coli* have no *a priori* interest in what the experimentalist wants. All strains and plasmids should be carefully maintained using sterile techniques. Passage of bacterial stocks should be minimized, and master stocks should always be prepared when an expression clone is first isolated or received. The expression system can, in many cases, be successfully stored as a plasmid-containing strain, frozen as a glycerol stock (containing 15% glycerol) at −70 °C. However, it is best to also store the components separately – the expression plasmid as a DNA preparation, ideally as an ethanol precipitate at −20 °C, and the *E. coli* host strain as a frozen glycerol stock at −70 °C. As has already been discussed, changes in the host, as well as in the plasmid, can lead to a decrease in the amount of recombinant protein produced. This problem can be reduced by producing a freshly transformed bacteria stock to start a large-scale fermentation, and this is the reason some people prefer to store plasmid DNA rather than *E. coli* expression strains. It is important to remember that freshly transformed colonies should be restreaked onto selective plates before growth in liquid culture; this avoids the small background of cells not carrying plasmids that are present on the original transformation plates and that can cause problems in liquid cultures. Cells lacking plasmids generally have a faster growth rate and can survive in liquid cultures containing plasmid-carrying cells that express enzymes that degrade the antibiotics. Contamination by cells lacking plasmids can significantly reduce the yield of recombinant proteins.

Even with these precautions, it is important to remember that the *E. coli* host can modify the plasmid. Wild-type *E. coli* contain a number of recombination systems that can act on plasmid DNA. This is a particular problem if a plasmid contains repeated sequences. Recombination between direct repeats is quite efficient in wild-type *E. coli*, but is greatly reduced in *recA* strains. Most of the *E. coli* hosts commonly used for producing recombinant proteins are *recA* deficient, and the use of such strains is strongly recommended.

Fermentation is an especially important part of protein expression. Using an identical strain and plasmid, slight alterations in growth conditions can make a substantial difference in the yield of the desired protein. Ideally, it is preferable to grow large amounts of *E. coli* that contain (relative to the host proteins) large amounts of the desired recombinant protein. In fermentation, the experimentalist controls the media, the temperature of fermentation and, in a large fermenter, the aeration and stirring. In rich media, if the culture is taken to saturation in shake flasks, it is usually possible to produce 4–8 g of *E. coli* (wet weight) per litre; substantially higher cell densities can be obtained in fermenters. The amount of *E. coli* that can be produced in actual practice and, more importantly, the amount of the recombinant protein relative to the *E. coli* host proteins, are sensitive to all of the variables. Unfortunately, there are relatively few hard and fast rules. To make matters worse, when the scale of the fermentation is changed, it is often necessary to develop new fermentation conditions; this is a particular problem when the scale is changed from shake flasks to a fermenter. Developing optimum conditions for the production of a recombinant protein in a fermenter usually requires repeated trials with the fermenter; this is both time consuming and expensive. Fortunately, with many expression systems, sufficient yields can be obtained using shake flasks, and, in cases where a fermenter is required, it is usually possible to get satisfactory (if suboptimal) results without extensive experimentation.

As a general rule, more total *E. coli* and more recombinant protein can be obtained by growing cells in rich media than in minimal media. The cells grow faster in such media, and inductions, in general, are fast and efficient. In some cases, it is necessary to choose between conditions that produce more total *E. coli* and conditions that produce a higher relative yield of the desired recombinant protein. Of these two, the relative yield of the desired protein is the more important. In designing fermentation protocols, it helps to understand how the host organism works. For example, *E. coli* are subject to catabolite repression. Given a choice of two carbon sources, *E. coli* will concentrate on the preferred carbon source to the exclusion of the second. *E. coli* prefer glucose to lactose; if a *lac*-based expression system is used, it is a good idea to avoid using growth media that contain glucose. Good results can often be obtained with media rich in amino acids (2 × YT or superbroth without glucose). In general, vigorous aeration is helpful. Begin fermentation trials by putting a relatively small amount of broth in a shake flask. For volumes of 1–1.5 l, aeration is much more efficient in wide-bottomed Fernbach flasks, and use of Fernbach flasks improves the yield of cells. In most fermenters, oxygen levels can be monitored, and air and O_2 delivery can be regulated to provide optimal levels of oxygen.

Finding an optimal temperature for maximum production of soluble recombinant protein usually requires experimentation. *E. coli* grow faster at 37 °C than at lower temperatures, and if a high-level expression of soluble protein is obtained under such conditions, there is rarely any advantage in looking further. However, in some cases, the relative yield of a recombinant protein can be substantially increased by growing *E. coli* expression strains at temperatures below 37 °C (discussed further below).

When screening expression constructs for production of recombinant protein, four scenarios are most commonly encountered:

(1) high-level expression of soluble recombinant protein;
(2) high-level expression of the recombinant protein, with a greater or lesser proportion of the protein in inclusion bodies;
(3) no expression or very low levels of expression; and
(4) lysis of cells.

The first result is usually the most welcome. Occasionally however, the expressed protein is smaller than predicted, presumably due to proteolysis. In such cases, production of a stable fragment suggests the presence of a compact, folded domain which might be worth pursuing for crystallography. However, it should be noted that not all soluble proteins are properly folded. Occasionally, misfolded proteins are expressed at high levels in soluble form. Such proteins usually exhibit aberrant behaviour during purification, such as aggregation or precipitation, migration as broad peaks during column chromatography and elution in the void volume during size-exclusion chromatography. In such cases, additional experimentation is required. Inclusion bodies are usually the result of improper protein folding, and cell lysis generally indicates severe toxicity. There are two obvious reasons for a failure to produce measurable amounts of a recombinant protein: either there is a problem at the level of transcription and/or translation, or there is proteolytic degradation of the protein. Some potential solutions to these problems are discussed below.

In some cases, the stability of the recombinant protein is related to its solubility. In general, only well folded proteins are soluble at high concentrations. In all living cells, protein concentrations are high; if a recombinant protein is expressed at a high level, it will be present inside the host cell at a high concentration. Protein folding is an active process in living cells. Molecular chaperones are used both to prevent unwanted interactions with other partially folded proteins and to promote the folding process. In some cases, when a recombinant protein is expressed at high levels, it will not fold properly in *E. coli*, either because it fails to interact properly with *E. coli* chaperones, or because it is made at such high levels that it overwhelms the available chaperones. In such cases the unfolded and/or partially folded protein may aggregate in inclusion bodies (Mitraki & King, 1989), which is both a blessing and a curse. Proteins in inclusion bodies are essentially immune to proteolytic degradation. Additionally, it is usually relatively easy to obtain the inclusion bodies in relatively pure form, making it simple to purify the recombinant protein. Unfortunately, the recombinant protein obtained from the inclusion bodies must be refolded. There are a variety of protocols for refolding proteins (discussed in Section 3.1.5.3), but few simple, universal prescriptions. Even under the most favourable conditions, with proteins that refold easily and (relatively) efficiently, the yield of properly folded material is often low. For some recombinant proteins obtained from inclusion bodies, it is the efficiency of the refolding step that limits the amount of material that can be obtained for crystallization. We will discuss this issue in more detail in Section 3.1.5.3.

The formation of inclusion bodies is the result of aggregation of non-native proteins. Factors that alter the folding pathway and/or affect the concentrations of unfolded or misfolded proteins can have a dramatic influence on the yield of soluble protein. It is not uncommon for recombinant proteins that form inclusion bodies when expressed at high levels (such as in a T7 expression system) to be present at undetectable levels when expressed at slightly lower levels (such as from a constitutive *lac* promoter). Presumably, in both cases the protein is failing to fold rapidly and efficiently. In the former case, the high levels of unfolded intermediates lead to the formation of inclusion bodies; in the latter case, the concentration of the unfolded protein is not sufficient to form inclusion bodies, and the unfolded protein is degraded. In some cases, it is relatively easy to express the protein, but variations in expression systems and/or culture conditions result in quite different yields of soluble and insoluble protein. In all of these situations, it is appropriate to try a number of different expression systems, with the hope that different kinetics of transcription and/or translation may result in concentrations of intermediates in which protein folding is favoured relative to aggregation and/or degradation. In some cases, reducing the temperature of fermentation is helpful (Schein & Noteborn, 1988). In addition to affecting rates of transcription and/or translation, temperature also affects folding. There are numerous examples in the literature where low temperature was essential to the recovery of soluble recombinant protein; however, there does not seem to be a general solution: optimal conditions seem to vary with each protein. In many cases, reducing the temperature of growth from 37 to 30 °C has improved the yield of soluble protein. However, temperatures as low as 17 °C have been reported as optimal for expression of some recombinant proteins (Biswas *et al.*, 1997), and there are anecdotal reports which indicate that protein can be successfully expressed as low as 14 °C. At low temperatures, growth of *E. coli* is quite slow. In most cases, inducible expression systems are used. Cells are grown to mid-logarithmic phase at 37 °C, and are cooled to the desired temperature just prior to induction (Yonemoto *et al.*, 1998). Significantly longer post-induction times are required for high protein yields, and soluble protein expression should be assessed over a 24-hour period to determine optimal times for maximum yields.

The rapid and proper folding of the overexpressed protein appears to be one of the most important factors in achieving high yields of recombinant proteins in *E. coli*. Attempts have been made to improve *in vivo* folding by co-expression of chaperones and other proteins that might aid the folding process (Wall & Plückthun, 1995; Cole, 1996; Georgiou & Valax, 1996). Once again, the usefulness of these strategies appears to be specific for individual recombinant proteins, although some folding components appear to be more broadly useful than others (Yasukawa *et al.*, 1995). A variety of proteins, including GroEL and GroES, DnaK and DnaJ, chaperones cloned from the host organism for the recombinant protein, thioridazine, protein disulfide isomerases (PDIs) and disulfide-forming protein DsbA, have all been used with varying degrees of success in different systems. It is unlikely that co-expression of chaperones or other proteins will be useful in overcoming folding or stability problems in proteins that are inherently unstable (those made unstable by removal of other domains, those lacking essential post-translational modifications or those failing to form essential disulfide bonds in the reducing environment inside *E. coli*).

Proteolytic degradation is an active process in *E. coli*, and several strategies for minimizing proteolysis of recombinant proteins have been developed (Enfors, 1992; Murby *et al.*, 1996). These strategies include secretion of proteins to the periplasm or external media, engineering of proteins to remove proteolytic cleavage sites, growth at low temperature and other strategies to promote folding, such as use of fusion proteins and co-expression with chaperones. One popular strategy, which unfortunately appears to be more protein-specific than might be expected, involves the use of *E. coli* strains that have genetic defects in the known proteolytic degradation pathways (Gottesman, 1990). If the desired protein is rapidly degraded in *E. coli*, and fermentation at lower temperatures does not solve the problem, *E. coli* deg$^-$ (degradation) mutants can be tried. However, the proteolytic machinery in *E. coli* is quite complicated, and a number of deg$^-$ mutants are available. All of the deg$^-$ mutants are more difficult to work with than wild-type strains, and there is no guarantee that expressing a particular recombinant protein in any of the available deg$^-$ mutants will cause a substantial increase in the yield of the recombinant protein. For these reasons, deg$^-$ mutants are usually tried only as a last resort. In most cases, proteolysis indicates a problem with protein folding, and efforts to improve protein folding are generally more fruitful than efforts to minimize proteolysis.

We briefly touched on the issue of the potential toxicity (to the *E. coli* host) of recombinant proteins when discussing constitutive and inducible vector systems. In general, the greatest difficulties are encountered with membrane proteins and enzymes. For the most part, enzymes are a problem because their enzymatic activities derange the host cell. For example, proteases are notoriously difficult to produce in large amounts. There are several ways to address this problem. First, as has already been discussed, it is important to use a tightly controlled inducible system if the recombinant protein is likely to disturb the metabolism of the *E. coli* host profoundly. If the recombinant protein is not properly folded, and is present primarily in inclusion bodies, the degree of toxicity is less, and often much less, than if the recombinant protein is present primarily in an active, soluble form. Although it is not the preferred procedure, and is not usually necessary, it is also possible to mutate the recombinant protein to reduce (or eliminate) its toxicity. In cases where the desired product is an enzyme, the enzyme can be inactivated by altering the amino acids at the active site.

Additional problems are encountered when trying to produce recombinant proteins that would, in higher eukaryotes, either be bound to or pass through membranes. There are several problems: *E. coli* do not usually grow well if they have large amounts of foreign protein in their membrane; this problem is compounded by the fact that the rules for membrane signals and signal processing are different in *E. coli* and higher eukaryotes. In general, the solution to this issue has been to express, in *E. coli*, only the internal or the external domain of membrane proteins from higher eukaryotes. Not only does this usually solve the problem of the toxicity of the protein in *E. coli*, but domains that are not directly associated with the membrane are usually much more soluble, easier to purify and much better candidates for crystallization. There is an additional issue. In contrast to the cell interior, which is, in general, a reducing environment, the milieu outside the cell is usually an oxidizing environment. Many of the proteins found on the outside of higher eukaryotic cells, or proteins that are exported from higher eukaryotic cells, have disulfide bridges that help stabilize their secondary and/or tertiary structure. Such disulfide bonds do not ordinarily form properly inside *E. coli*, and it can be much more difficult to obtain recombinant proteins that have extensive and complex disulfide bridges in a properly folded form from *E. coli*.

3.1.4.2. *Yeast*

Yeasts are simple eukaryotic cells. Considerable effort has been expended in studying brewers' yeast, *Saccharomyces cerevisiae*, and in developing plasmid systems and expression vectors that can be used in this organism. Recently, methylotrophic yeasts, most notably *Pichia pastoris*, have been developed as alternative systems that offer several advantages over *S. cerevisiae*. Although yeast expression systems are reasonably robust, the expertise required to use these systems effectively is not as widespread as the corresponding expertise for the manipulation of *E. coli* strains. Nor are the tools, media and reagents necessary to grow yeast and select for the presence of expression plasmids as broadly available as those used for *E. coli* systems. However, the increasing commercial availability of complete kits (such as *Pichia* expression systems from Invitrogen) is making yeast systems more accessible.

While yeast systems do offer some advantages relative to *E. coli*, these advantages are, in general, modest. One primary advantage, the ability to produce large amounts of biomass using simple, inexpensive culture media, is probably more important for industrial-scale protein expression than for most laboratory applications, even those involving crystallography, which requires more protein than most simple biochemical experiments. Yeast systems do not, in general, offer solutions to some of the most difficult problems encountered when trying to express recombinant proteins in *E. coli.* Specifically, the problem of mimicking the post-translational modifications found in higher eukaryotes (particularly glycosylation), which has not been solved for *E. coli*, has not been solved in yeast either. None of the available systems recapitulates the post-translational modifications found in higher eukaryotes. Additionally, yeast systems introduce some new problems not seen with *E. coli* expression systems, specifically genetic instability and hyperglycosylation, both of which are more problematic in *S. cerevisiae* than in *Pichia*.

Yeast systems are perhaps most valued for high-level production of secreted proteins. For some naturally secreted proteins, passage through the secretory pathway is necessary for proteolytic maturation, glycosylation and/or disulfide bond formation and is essential for proper folding or function. But secretion is complex, and numerous factors, such as the signal sequence, gene copy number and host strain, can be critical for high-level expression. Secretion can significantly simplify purification, since secreted recombinant proteins can constitute as much as 80% of the protein in the culture medium. However, degradation of secreted proteins can be a major problem. In some instances, proteolysis has been minimized by alteration of the pH of the culture medium, by addition of amino acids and peptides, and by use of protease-deficient strains (Cregg *et al.*, 1993).

The rules for expression of proteins in yeast are not the same as those used either in *E. coli* or in higher eukaryotes. In yeast, as in *E. coli*, cDNA sequences from a higher eukaryote must be tailored for high-level expression, following rules that are fairly well understood. Yeast grows at 25–30 °C and has a slower growth rate than *E. coli* (under typical growth conditions, yeast has a doubling time of approximately 90 min, compared to 30 min for *E. coli*). Transformation of yeast can be achieved using competent cells, sphaeroplasts or electroporation, but by any technique it is less efficient than the transformation of *E. coli*. For these reasons, most yeast plasmids are designed to replicate both in *E. coli* and yeast; the DNA manipulations are done using an *E. coli* host, and the completed expression plasmid is introduced into yeast as the final step in the process.

Most expression vectors in *S. cerevisiae* are based on the yeast 2μ plasmid (Beggs, 1978; Broach, 1983) that is maintained as an episome, present at approximately 100 copies per cell. Plasmid instability can result in loss of expression during production, and integrating vectors have been developed that provide greater stability, albeit with levels of expression that are, in general, lower than the plasmid systems. Both constitutive and tightly regulated inducible expression systems have been developed using a variety of promoters. The most widely used systems involve galactose-regulated promoters, such as *GAL1*, which are capable of rapid and high-level induction. An extensive review of recombinant gene expression in yeast (Romanos *et al.*, 1992) is highly recommended as a resource for anyone seriously contemplating the expression of recombinant proteins in *S. cerevisiae.*

In terms of high-level expression, the *Pichia* system may ultimately prove to be more useful than *S. cerevisiae* (for reviews see Cregg *et al.*, 1993; Romanos, 1995; Hollenberg & Gellissen, 1997). There is considerable interest in developing the *Pichia* system for the expression of recombinant proteins, especially for industrial applications, and there has been sufficient progress made to support the publication of a useful monograph for specific techniques (Higgins & Cregg, 1998). *Pichia* offer several advantages over *S. cerevisiae*. Intracellular protein expression can be extremely high in *Pichia*, reaching grams per litre of cell culture. Large amounts of secreted proteins can be produced using media that are almost protein-free, although the expression levels are not quite as high as for intracellular proteins. *Pichia* can be cultured to very high cell density with good genetic stability. Additionally, hyperglycosylation is less of a problem in *Pichia*, which typically have shorter outer-chain mannose units (less than 30 outer-chain residues) than *S. cerevisiae* (greater than 50 residues) (Grinna & Tschopp, 1989).

Methylotrophic yeasts, which are able to use methanol as their sole carbon source, contain regulated methanol enzymes that can be induced to give extremely high levels of expression. In *Pichia* expression systems, the gene that encodes alcohol oxidase (*AOX1*) is most commonly used for the expression of foreign genes, but constitutive promoters are also available. Heterologous genes are inserted into vectors and then integrated into the *Pichia* genome, either duplicating or replacing (transplacement) the target gene, depending on how the linearized vector is constructed. High-level expression relies on integration of multiple copies of the foreign gene and, since this varies significantly, screening colonies to obtain clones with the highest levels of expression is required. Culture conditions and induction protocols are critical for optimal expression. Since *Pichia* are readily oxygen-limited in shake flasks, growth in fermenters is required for high-level expression (approximately five- to tenfold greater than in shake flasks).

Numerous factors make yeast expression systems significantly less straightforward than those of *E. coli*. In addition to the considerations mentioned above, it should be noted that yeast cells are surrounded by a tough cell wall and are therefore notoriously difficult to break. This makes the problem of purification of intracellular protein from yeast that much more difficult. Given the many complexities of expression in yeast, it is usually better to begin with an *E. coli* expression system and move to yeast only if the results obtained with *E. coli* systems are unacceptable. If yeast is used as an expression system, careful attention should be paid to maintaining defined stocks of the expression strain and the corresponding expression plasmids. Despite the availability of comprehensive kits, if the researcher does not have considerable experience with yeast, the enlistment of an experienced colleague is recommended.

3.1.4.3. Baculoviruses and insect cells

Baculovirus expression systems are becoming increasingly important tools for the production of recombinant proteins for X-ray crystallography. The insect cell–virus expression systems are more experimentally demanding than bacteria or yeast, but they offer several advantages. Expression of some mammalian proteins has been achieved in baculovirus when simpler expression systems have failed. Because insects are higher eukaryotes, many of the difficulties associated with expression of proteins from higher eukaryotes in *E. coli* do not apply: there is no need for a Shine–Dalgarno sequence, no major problems with codon usage and fewer problems with a lack of appropriate chaperones. Although glycosylation is not the same in insect and mammalian cells, in some cases it is close enough to be acceptable. In addition, for many crystallography projects, minimizing glycosylation is helpful, so that it may be more appropriate to modify the gene or protein to avoid glycosylation (or minimize it) than to try to find ways to recapitulate the glycosylation pattern found in mammalian cells. As is the general case in biotechnology, the development of baculovirus expression systems is work in progress. Progress has been made towards making recombinant proteins in insect cells with glycosylation patterns that match those in mammalian cells (reviewed by Jarvis *et al.*, 1998). Baculovirus systems allow expression of recombinant proteins at reasonable levels, typically ranging from 1–500 mg l^{-1} of cell culture. Considerable work has gone into the development of convenient transfer vectors, and baculovirus expression kits are available from more than ten different commercial sources.

Baculoviruses usually infect insects; in terms of the expression of foreign proteins, the important baculoviruses are the *Autographa californica* nuclear polyhedrosis virus (AcNPV) and the *Bombyx mori* nuclear polyhedrosis virus (BmNPV). AcNPV has been used more widely than BmNPV in cell-culture systems; the BmNPV virus is used primarily to express recombinant proteins in insect larvae. The advantage of BmNPV is that it grows well in larger insect larvae, making the task of harvesting the haemolymph easier. Proteins expressed for crystallography have all been, to the best of our knowledge, expressed using the AcNPV virus system; we will not discuss the BmNPV virus expression system here. Anyone wishing to learn more about either AcNPV or BmNPV is urged to consult two useful monographs: *Baculovirus Expression Protocols* (Richardson, 1995) and *Baculovirus Expression Vectors: A Laboratory Manual* (O'Reilly *et al.*, 1992). There are also shorter reviews that are quite helpful (Jones & Morikawa, 1996; Merrington *et al.*, 1997; Possee, 1997).

In nature, in the late stage of replication in insect larvae, nuclear polyhedrosis viruses produce an occluded form, in which the virions are encased in a crystalline protein matrix, polyhedrin. After the virus is released from the insect larvae, this proteinaceous coat protects the virus from the environment and is necessary for the propagation of the virus in its natural state. However, replication of the virus in cell culture does not require the formation of occlusion bodies. In tissue culture, the production of occlusion bodies is dispensable, and the primary protein, polyhedrin, is not required for replication. Cultured cells infected with wild-type AcNPV produce large amounts of polyhedrin; cells infected with modified AcNPV vectors, with other genes inserted in place of the polyhedrin gene (or in place of another highly expressed gene, *p10*, that is dispensable in cultured cells), can express impressive amounts of the recombinant protein.

The AcNVP genome is 128 kb, which is too large for convenient direct manipulations. In most cases, novel genes are put into the AcNPV genome by homologous recombination using transfer vectors. Transfer vectors are small bacterial plasmids that contain AcNPV sequences that allow homologous recombination to direct the insertion of the transfer vector into the desired place in the AcNPV genome (often, but not always, the polyhedrin gene). Originally, the purified circular DNA from AcNPV and the appropriate transfer plasmids were simply cotransfected onto monolayers of insect cells. Plaques develop, and if the insertion is targeted to the polyhedrin gene, plaques that contain viruses that retain the ability to make polyhedrin (those that contain the wild-type virus) can be distinguished in the microscope from plaques that do not. This technique works, but has been largely replaced by systems that make it easier to obtain and/or find the recombinant plaques. The AcNPV genome is circular; if the DNA is linearized, it will not produce a replicating virus unless the break is repaired. The repair process is facilitated by the presence of homologous DNA flanking the break. Systems have been set up to exploit this property to increase the efficiency of the generation of vectors that carry the desired insert. Basically, the genome of the AcNPV vector is modified so that there is a unique restriction site at the site where the transfer vector would insert. Linear AcNPV DNA is cotransfected with a transfer vector. This can produce stocks in which greater than 90% of the virus is recombinant. Systems have also been developed in which a DNA insert can be ligated directly into a linearized AcNPV genome. This protocol also produces a high yield of recombinant virus (Lu & Miller, 1996).

There are also a number of systems that allow either the selection or, more often, the ready identification of recombinant virus. The marker most commonly used for this purpose is β-galactosidase; a number of AcNPV vectors or transfer systems that make use of β-galactosidase are commercially available. Once a recombinant plaque is identified, it should be purified through multiple rounds of plaque purification to ensure that a homogeneous stock has been prepared. Several independent isolates should be prepared and each checked for expression of the desired protein.

There are several important things to consider when setting up the cell-culture system. Although most baculoviruses have a relatively restricted host range, and AcNPV was first isolated from alfalfa looper (*Autographa californica*), for the purpose of expressing foreign proteins, it is usually grown in cells of the fall armyworm (*Spodoptera frugiperda*) or the cabbage looper (*Trichoplusia ni*). The isolation and purification of the appropriate AcNPV vectors are usually done in monolayer cultures. In contrast, the production of large amounts of recombinant protein is usually done in suspension cultures. There is also the issue of whether or not to include fetal calf serum in the culture media. In theory, since the cells can be grown in serum-free media, which saves money and makes the subsequent purification of the recombinant protein simpler, serum-free culture is the appropriate choice. However, growing cells in serum-free media is a trickier proposition, and the cells are more sensitive to minor contaminants. As a general rule, high-level production of recombinant proteins using a baculovirus vector requires host cells that are growing rapidly; this is sometimes easier to achieve with serum-containing media. It is not always a simple matter to switch cells adapted to growth on plates to suspension culture, nor is it always easy to switch cells grown in the presence of serum to serum-free culture. Since the vector is a virus, it is usually more convenient to use cells adapted to different conditions than to try to adapt the cells. However, the relative yield of the recombinant protein will not necessarily be the same in different cells grown under different culture conditions.

Although baculoviruses, particularly AcNPV, are convenient vectors, the expression of the recombinant protein is carried out by the insect cell host. Baculovirus infection kills the host cell, so it is not possible to use baculoviruses to derive insect cell cultures that continuously express a recombinant protein. It is possible, however, to introduce DNA segments directly into insect cells and derive cell lines that stably express a recombinant protein; there are constitutive and inducible promoters that can be used in insect cell systems (McCarroll & King, 1997; Pfeifer, 1998). Basically, the protocols used to introduce DNA expression constructs into cultured insect cells are similar to those used in cultured mammalian cells ($CaPO_4$, electroporation, liposomes *etc.*), and similar selective protocols are used (G418, hygromycin, puromycin *etc.*).

Expression systems have been prepared based on baculovirus immediate early promoters and on cellular promoters, including the hsp70 promoter and metallothionein (McCarroll & King, 1997; Kwong *et al.*, 1998; Pfeifer, 1998). Insect cells are, in general, easier (and cheaper) to grow in culture than mammalian cells, although many of the problems that exist in mammalian cell culture also exist in insect cell culture. Relative to the baculovirus system, the use of stable insect cell lines not only allows the continuous culture of cells that contain the desired expression system (provided the expressed protein is not too toxic), it also permits the use of *Drosophila* cell lines, which appear to have some advantages for the high-level production of recombinant proteins.

Compared to bacteria or yeast cells, cells from higher eukaryotes are quite delicate, and considerable care must be taken in cell culture. The cells are subject to shear stress, which can be a problem in stirred and/or shaken cultures; some researchers use airlift fermenters to help alleviate the problem. Compared to yeast and bacterial cells, cultured cells grow relatively slowly and require rich media that will support the rapid growth of a wide variety of unwanted organisms, so special care must be taken to avoid contaminating the cultures. Antibiotics are commonly used; however, antibiotics will not, in general, prevent contamination with yeasts or moulds, which often cause the greatest problems. If the baculovirus system is used, then the cells and viruses are kept separate, and the cells are relatively standard reagents. If there is contamination, the contaminated cultures can be discarded and replaced with fresh cells (and viruses). Stable transformed insect cells that express a recombinant protein must be kept free of all contaminants. As is always the case, both cells and viruses should be carefully stored. Any useful recombinant baculovirus can be easily stored as DNA.

3.1.4.4. Mammalian cells

In some cases, however, even the baculovirus and/or insect cell expression systems are not able to make the desired recombinant protein product. If the recombinant protein is sufficiently important, it can be produced in cultured mammalian cells. Although biotechnology companies have demonstrated that it is feasible to produce kilograms of pure recombinant proteins using cultured mammalian cells, the effort required to produce tissue culture cells that express high levels of recombinant protein is substantial, and the costs of growing large amounts of tissue culture cells are beyond the means of all but the best-funded laboratories. To make matters worse, there are no well defined plasmids that reliably and stably replicate in mammalian cells. It may be possible to develop reliable episomal replication systems based on viral replicons; however, even the best developed viral episomes are still not entirely satisfactory (see, for example, Sclimenti & Calos, 1998). Cell lines are usually prepared by transfection; following transfection, some of the cells (usually a small percentage) will incorporate transfected DNA into their genomes. A number of agents can be used to transfect DNA; these include, but are not limited to, $CaPO_4$, DEAE Dextran, cationic lipids *etc.* This is a complex and poorly defined process; the transfected DNA is often incorporated into complex tandem arrays. Neither the amount of transfected DNA nor its location in the host genome is controlled in a standard transfection; as a consequence, the expression level varies substantially from one transfected cell to another. This makes the process of creating mammalian cell lines that efficiently and stably express a recombinant protein a labour-intensive process. Ordinarily, the DNA segment carrying the gene for the desired recombinant protein is linked to a selectable marker; selection for the marker is usually sufficient to cause the retention of the gene for the desired recombinant protein, provided that it is not toxic to the host cell. The tandem arrays produced by transfecting DNA into mammalian cells are often unstable. Recombination within the tandem array can decrease (or less commonly increase) the number of copies of the transfected gene. It is possible to take advantage of this instability. Selection protocols, which usually involve the DHFR gene and methotrexate, have been developed that can select for cells that have the DNA segments containing both the selectable marker and the gene for the desired recombinant protein in higher copy number (Kaufman, 1990); these can be used to develop cell lines that express high levels of the recombinant protein.

There are alternative methods that can be used to deliver an expression construct to a cultured mammalian cell. For example, the DNA can be introduced by electroporation, and homologous recombination can be used to embed an expression construct at a specific place in the host genome. However, such strategies, while in some ways more elegant than simple DNA transfection, do not appear to simplify the problem of creating a cell line that produces large amounts of a specific gene product. There are also a variety of viral vectors that can be used to introduce genes into cells either transiently or stably. At the time of writing, viral vector systems, which are extremely useful for studying the effects of expressing foreign genes in cultured mammalian cells, do not appear to offer any obvious advantages for the preparation of cultured cells that can express the relatively large amounts of recombinant protein needed for crystallography. However, this is an area where the research effort is particularly intense, so it is entirely possible that in the near future there will be a viral vector (or vectors) which will offer significant advantages for inducing high-level expression of recombinant proteins in cultured mammalian cells.

Until relatively recently, one of the primary problems in working with expression systems in cultured mammalian cells has been the lack of a tightly regulated inducible system. This has made the high-level expression of proteins that are deleterious to the growth of the cell an exceptionally difficult problem. The promoters originally used for inducible expression in cultured mammalian cells (metallothionein, glucocorticoid responsive *etc.*) tend to be leaky in the absence of the inducer. If cell lines were chosen in which the desired protein was not synthesized in the absence of the inducer, the level of the recombinant protein that could be made in the presence of the inducer was usually, but not always, low.

There has been progress in the development of more efficient and reliable inducible promoters for cultured mammalian cells. These systems are complex and require cell lines that express regulatory proteins not normally found in cultured mammalian cells. In this sense they are the logical counterparts of the T7 RNA polymerase/*lac* expression systems for *E. coli* already discussed in this chapter. The best developed of the engineered systems designed to permit the inducible expression of genes in mammalian cells are (1) the tetracycline system, (2) the F506/rapamycin system, (3) the RU486 system and (4) the ecdysone system (Saez *et al.*, 1997; Rossi & Blau, 1998).

Although these four inducible systems differ in important ways, there are common themes. Firstly, in all cases, the small molecule used as the inducer is not normally a regulator of gene expression in mammalian cells. This means that application of the inducer to cells should not substantially perturb the normal pattern of gene expression and, by implication, the health of the cells. Secondly, the DNA target sequences used to activate the expression of the recombinant gene/protein are not sequences known to be associated with the expression of normal cellular genes. This should also help prevent the activation of normal cellular genes when these systems are used.

In all of these systems, the specific regulation of an introduced gene requires a special regulatory protein that interacts with the appropriate small-molecule inducer and recognizes the requisite DNA target sequence that is linked to the gene of interest. These regulatory proteins, which were derived, at least in part, from regulatory proteins from nonmammalian hosts, must be present in the cell line for induction/regulation to occur. This means that either the researcher must choose from a relatively limited set of cells that already express the desired regulatory factor or face the problem of introducing (and carefully monitoring the proper expression and function of) both the regulatory factor and the desired recombinant protein. Considerable effort has been put into the development of each of these systems and significant progress has been made. At the moment, the tetracycline inducible system is probably the most fully developed; however, this is a fast moving area of research, and it is not now certain which of these systems will ultimately prove to be the most useful for the high-level expression of recombinant proteins in cultured mammalian cells.

Suffice it to say, however, that despite all the efforts of a large group of talented researchers, the systems available for use in cultured mammalian cells are much less well defined and much more difficult to use than the corresponding *E. coli* and yeast expression systems, and anyone who is not well versed in the problems associated with using expression systems designed for cultured mammalian cells should be most cautious about using them for the large-scale production of recombinant protein.

Despite these problems, mammalian (and, less frequently, insect cell) expression systems have been used to prepare proteins for crystallography. For example, in the recent determination of the X-ray structure of a complex between a portion of CD4, a modified version of HIV-1 gp120 and the Fab fragment of a monoclonal antibody, each of the proteins was made in cultured cells, but three different types of cultured cells were used. The two-domain segment of CD4 was made in Chinese hamster ovary cells. The monoclonal antibody used to prepare the Fab was made in an immortalized human B cell clone, and the core of gp120 in *Drosophila* Schneider 2 cells under the control of a metallothionein promoter (Kwong *et al.*, 1998).

Tissue culture cells are much more difficult to grow than either yeast or *E. coli.* As has already been discussed in Section 3.1.4.3, there is the issue of using calf (or fetal calf) serum. A relatively small number of mammalian cell lines have been developed that will grow on defined media without serum; this is an advantage, but the media are still relatively costly. Mammalian cell lines expressing recombinant proteins must be maintained for long periods under carefully controlled conditions, both to ensure that the expression of the recombinant protein is maintained and to avoid contamination of the cultures with bacteria, yeast or moulds. Because the cells grow relatively slowly (doubling times are commonly 24–48 hours), it is usually not a simple task to produce 10–20 g (wet weight) of cells – something that can be done overnight with *E. coli.* If a useful cell line is obtained, it should be carefully stored in multiple aliquots. Cultured cells are routinely stored (in the presence of cryoprotectants) in liquid nitrogen. Short-term storage at −70 °C is an acceptable practice; however, long-term storage will be much more successful if lower temperatures are used.

3.1.5. Protein purification

3.1.5.1. Conventional protein purification

Those of us old enough to remember the task of purifying proteins from their natural sources, using conventional (as opposed to affinity) chromatography, where a 5000-fold purification was not unusual and the purifications routinely began with kilogram quantities (wet weight) of *E. coli* paste or calves' liver, are most grateful to those who developed efficient systems to express recombinant proteins. In most cases, it is possible to develop expression systems that limit the required purification to, at most, 20- to 50-fold, which vastly simplifies the purification procedure and concomitantly reduces the amount of starting material required to produce the 5–10 mg of pure protein needed to begin crystallization trials. This does not mean, however, that the process of purifying recombinant proteins is trivial. Fortunately, advances in chromatography media and instrumentation have improved both the speed and ease of protein purification. A wide variety of chromatography media (and prepacked columns) are commercially available, along with technical bulletins that provide detailed recommended protocols for their use. Purification systems (such as Pharmacia's FPLC and ÄKTA systems, PerSeptive Biosystems' BioCAD workstations and BioRad's BioLogic systems) include instruments for sample application, pumps for solvent delivery, columns, sample detection, fraction collection and information storage and output into a single integrated system, but such systems are relatively expensive. Several types of high capacity, high flow rate chromatography media and columns (for example, Pharmacia's HiTrap products and PerSeptive Biosystems' POROS Perfusion Chromatography products) have been developed and are marketed for use with these systems. However, the use of these media is not restricted to the integrated systems; they can be used effectively in conventional chromatography without the need for expensive instrumentation.

In designing a purification protocol, it is critically important that careful thought be given to the design of the protocol and to a proper ordering of the purification steps. In most cases, indi-

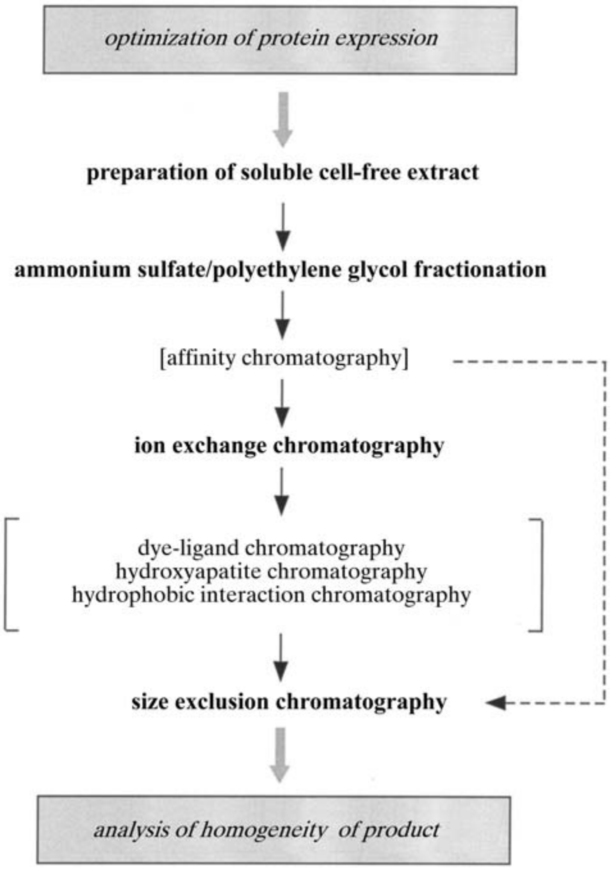

optimization of protein expression
preparation of soluble cell-free extract
ammonium sulfate/polyethylene glycol fractionation
[affinity chromatography]
ion exchange chromatography
dye-ligand chromatography
hydroxyapatite chromatography
hydrophobic interaction chromatography
size exclusion chromatography
analysis of homogeneity of product

Figure 3.1.5.1
Protein purification strategy. Purification of proteins expressed at reasonably high levels typically requires only a limited number of chromatographic steps. Additional chromatography columns (indicated in brackets) can be included as necessary. Affinity chromatography can allow efficient purification of fusion proteins or proteins with well defined ligand-binding domains.

vidual purification steps are worked out on a relatively small scale, and an overall purification scheme is developed based on an ordering of these independently developed steps. However, the experimentalist, in planning a purification scheme, should keep the amount of protein needed for the project firmly in mind. In general, crystallography takes a good deal more purified protein than conventional biochemical analyses. Scaling up a purification scheme is an art; however, it should be clear that purification steps that can be conveniently done in batch mode (precipitation steps) should be the earliest steps in a large-scale purification, chromatographic steps that involve the absorption and desorption of the protein from columns (ion-exchange, hydroxyapatite, hydrophobic interaction, dye-ligand and affinity chromatography) should be done as intermediate steps, and size exclusion, which requires the largest column volumes relative to the amount of protein to be purified, should generally be used only as the last step of purification. If reasonably good levels of expression can be achieved, most recombinant proteins can be purified using a relatively simple combination of the previously mentioned procedures (Fig. 3.1.5.1), requiring a limited number of column chromatography steps (generally two or three).

All protein purification steps are based on the fact that the biochemical properties of proteins differ: proteins are different sizes, have different surface charges and different hydrophobicity. With the exception of a small number of cases involving proteins that have unusual solubility characteristics, batch precipitation steps usually do not provide substantial increases in purity. However, precipitation is often used as the first step in a purification procedure, in part because it can be used to separate protein from nucleic acids. Nucleic acids are highly charged polyanions; the presence of nucleic acid in a protein extract can dramatically decrease the efficiency of column chromatography, for example by saturation of anion-exchange resins. If the desired protein binds to nucleic acids and the nucleic acids are not removed, ion-exchange chromatography can be compromised by the interactions of the protein and the nucleic acid and by the interactions of the nucleic acid and the column. The most commonly used precipitation reagents are ammonium sulfate and polyethylene glycols. With little effort, the defined range of these reagents needed to precipitate the protein of interest can be determined. However, if the precipitation range is broad, it may be only marginally less efficient simply to precipitate the majority of proteins by addition of ammonium sulfate to 85% saturation or 30% polyethylene glycol 6000. Precipitation can be a useful method for concentrating proteins at various steps during purification and for storing proteins that are unstable upon freezing or upon storage in solution.

Column chromatography steps in which the protein is absorbed onto the resin under one set of conditions and then eluted from the column under a different set of conditions can produce significant purification. Anion-exchange chromatography is usually a good starting point. Most proteins have acidic pIs, and conditions can often be found that allow binding of the protein to anion-exchange matrices. Elution of the protein in an optimized gradient often yields greater than tenfold purification. If conditions cannot be found under which the protein binds to an anion-exchange resin, a reverse strategy can be advantageous. Conditions can be adjusted to promote the binding of most proteins, yielding a flow-through fraction enriched for the protein of interest. Fewer proteins interact with cation-exchange resins; if the desired protein binds, this can be a powerful step. Use of an anion exchanger does not necessarily preclude use of a cation-exchange column; under appropriately chosen sets of conditions (most notably adjustment of pH), a single protein can bind to both resins. Hydroxyapatite resins provide a variation of ion-exchange chromatography that can be extremely powerful for some proteins. While hydroxyapatite columns (traditionally just a modified form of crystalline calcium phosphate) have the reputation of slow flow rates, alternative matrices exhibiting improved flow properties have made hydroxyapatite chromatography significantly less tedious. Hydrophobic interaction chromatography can also provide significant purification and has the advantage that the protein is loaded onto the resin in a high ionic strength buffer, making it a good step following ammonium sulfate precipitation. Proteins can behave very differently with different hydrophobic matrices, and an exploration of a variety of different resins is often a worthwhile exercise. Several tester kits containing an assortment of resins are commercially available. Dye-ligand chromatography can also be explored using an assortment of test columns. Several of the dyes, most notably Cibacron Blue F3GA, have structures that resemble nucleotides and have been useful in purifying kinases, polymerases and other nucleotide-binding proteins. However, many proteins have significant affinity for various dyes, independent of nucleotide-binding activity, and the usefulness of dye-ligand chromatography for any specific protein needs to be determined empirically.

Size-exclusion chromatography, which does not involve absorption of the protein onto the matrix, rarely provides as much purification as the chromatography steps described above.

However, this can be a good step to include at the end of a purification scheme. Isolation of a well defined peak in the included volume separates intact, properly folded protein from any damaged/aggregated species that may have been generated during the purification procedure. Furthermore, size-exclusion chromatography can provide a useful indication of whether the protein is a well defined, folded, compact, monodisperse population, or whether it is oligomerizing, aggregating or exists in an unfolded or extended form. Although size-exclusion chromatography does not provide a definitive analysis of such behaviour, migration of the protein consistent with its expected molecular weight is generally a good sign; elution of a relatively small protein in the void volume suggests a need for further analysis. Size-exclusion-chromatography media are available for the fractionation of proteins in many different size ranges. Substantial improvement in purification can be achieved by choosing a size range that is optimal for the protein of interest. However, the ability of size-exclusion columns to separate proteins of different molecular weights is dependent on the amount of protein loaded on to the column. Better purification is obtained when relatively small volumes of protein (generally 1–2% of the column bed volume) are loaded on size-exclusion columns. If really large amounts of protein are needed for a crystallography project, it can be difficult (and expensive) to set up size-exclusion columns large enough to fractionate the desired amount of protein.

3.1.5.2. Affinity purification

The most powerful purification steps are those that most clearly differentiate the desired protein from the other proteins present. Many proteins bind specifically to substrates, products and/or other proteins. In some cases, it is possible to use specific ligands to design columns to which the desired protein will bind selectively. For example, it may be possible to chemically link the substrate or product of a particular enzyme to an inert support. If the modification to the small molecule needed to link it to the support is chosen so that it does not interfere with the binding of the enzyme, the modified resin can be used to purify the protein by affinity chromatography. If, as expected, the desired protein binds selectively, it can usually be eluted by washing the column with the same substrate used to prepare the column. This is a powerful procedure and can produce greater than 100-fold purification in a single step. Although this is a fairly well developed field, and there is sufficient experience to show that the process is often fruitful, it must be said that the development of an efficient and effective affinity column and an attendant purification procedure can be long, difficult and, depending on the ligand and/or activated resin, sometimes expensive. In addition, the preparation of the column usually involves some moderately sophisticated chemistry; if such a step is contemplated, it is helpful to have the requisite chemical sophistication.

Immuno-affinity chromatography is a classic affinity method that uses affinity media created by coupling antibodies (either monoclonal or polyclonal) specific for the protein of interest to an activated resin. Theoretically, if good antibodies are available in sufficient quantity, this should be a powerful and widely applicable method. However, immuno-affinity chromatography has two severe limitations. In most cases, the interaction between the antibody and antigen is so tight that harsh conditions are necessary to elute the bound protein, potentially resulting in denaturation of the protein. Additionally, scaling up the procedure for isolation of 5–10 mg of protein is usually not feasible because of the large quantities of antibody required for column preparation.

Because the process of affinity chromatography is so powerful, and the development of a specific affinity column is difficult, considerable effort has been expended on the development of general procedures for affinity chromatography. As discussed previously, it is possible to modify the recombinant protein so that it contains a sequence element that can be used for affinity chromatography. Numerous systems are being marketed that pair vectors for creation of fusion proteins with appropriate resins for affinity purification. Examples of these fusion element–affinity resin pairs include His_6–Ni^{2+}-nitrilotriacetic acid, biotinylation-based epitopes–avidin, calmodulin-binding peptide–calmodulin, cellulose or chitin-binding domains–cellulose or chitin, glutathione S-transferase–glutathione, maltose-binding domain–amylose, protein A domains–IgG, ribonuclease A S-peptide–S-protein, streptavidin-binding peptides–streptavidin and thioredoxin–phenylarsine oxide.

Several considerations are important in choosing a strategy for expression and purification of a fusion protein. Some of these issues have already been discussed (see Section 3.1.3.3). The most fundamental, and unfortunately least predictable, is what construct will produce large amounts of the recombinant protein. The presence of fusion proteins and/or purification tags perturbs the recombinant protein to a greater or lesser degree. Perturbation can in some cases be beneficial, with the fusion protein aiding *in vivo* folding or *in vitro* refolding. There is also the issue of whether or not to remove the tag or fusion protein. Removal of the tag usually involves engineering a site for a specific protease, digestion with that protease and subsequent purification to isolate the final cleaved product. Additional issues should also be addressed. Most of the well developed systems allow for the elution of the fusion protein from the affinity resin under relatively mild conditions that should not harm most proteins. However, the method of elution should be considered with respect to the specific requirements of the protein of interest. Since the costs of using the different systems on a large scale varies significantly, it is wise to calculate the expense associated with scaling up, allowing for the cost and lifetime of the affinity resin, the cost of the reagent used for elution and the cost of the protease if the tag is to be removed. Finally, the nature of the fusion element–affinity resin interaction should be considered. Some of these systems, such as the His_6 tag, can be used for purification under denaturing conditions, which is a considerable advantage if the desired recombinant protein is found in inclusion bodies.

3.1.5.3. Purifying and refolding denatured proteins

As we have already discussed, expressing high levels of recombinant prokaryotic or eukaryotic proteins in *E. coli* can lead to the production of improperly folded material that aggregates to form insoluble inclusion bodies (Marston, 1986; Krueger *et al.*, 1989; Mitraki & King, 1989; Hockney, 1994). Inclusion bodies can usually be recovered relatively easily, following lysis of cells by low-speed centrifugation (5 min at 12 000 *g*); inclusion bodies are larger than most macromolecular structures found in *E. coli* and denser than *E. coli* membranes. Care should be taken to achieve complete lysis, since an intact bacterial cell that remains after lysis will co-sediment with the inclusion bodies. In most (but not all) cases, the inclusion bodies contain the desired recombinant protein in relatively pure form.

In such cases, the problem lies not with the purification of the protein, but in finding a proper way to refold it.

Various general procedures for refolding proteins from inclusion bodies have been described (Fischer *et al.*, 1993; Werner *et al.*, 1994; Hofmann *et al.*, 1995; Guise *et al.*, 1996; De Bernardez Clark, 1998), and the literature is filled with examples of specific protocols. The insoluble inclusion bodies are usually solubilized in a powerful chaotropic agent like guanidine hydrochloride or urea. In general, detergents are not recommended. The denaturant is sequentially removed by dilution, dialysis or filtration. Both rapid dilution and slow removal of the denaturant have been used successfully. In most refolding protocols, relatively dilute solutions of the protein are used to avoid protein–protein interactions, and, if necessary, glutathione or some other thiol reagent is included in the buffer to accelerate correct pairing of disulfides. After a refolding procedure, the properly folded soluble protein must be separated from the fraction that did not fold appropriately. Improperly refolded proteins are relatively insoluble and can usually be removed by centrifugation. It is sometimes profitable to try to refold the recovered insoluble material a second time.

Once soluble protein has been obtained, conventional purification procedures may be employed. It should be noted that recovery of soluble protein is not necessarily an indication that the protein exists in a native state. Quantitative assays of protein activity should be used to characterize the protein, if such assays exist. Alternatively, the behaviour of the refolded protein should be critically assessed during subsequent purification steps; an improperly folded protein will be prone to aggregation, will generally give broad and/or trailing peaks during column chromatography and will migrate faster than expected during size-exclusion chromatography. Some proteins are more amenable to refolding than others. As has already been pointed out, if a protein has a complex array of disulfide bonds, it is usually more difficult to refold than a protein without disulfide bonds. Greater success in refolding is generally obtained with proteins composed of single domains than with multidomain proteins.

3.1.6. Characterization of the purified product

3.1.6.1. Assessment of sample homogeneity

The ultimate test of the usefulness of a purified protein for crystallization is determined by the actual crystallization trials. However, before such trials begin, the properties and purity of the recombinant protein should be carefully checked. There is some disagreement about the degree of purity required for crystallization. In the earliest days of protein purification, crystallization was used as a technique for the purification of proteins, and it is clear that absolute purity is not a requirement for the preparation of useful protein crystals. However, most practitioners of the art of crystallization prefer to use highly purified proteins for crystallization trials. There are several reasons for this. It is easier to achieve the high concentrations of protein (greater than 10 mg ml^{-1}) usually needed for crystallization if the protein is pure, and the behaviour of highly purified proteins is more reproducible. A homogeneous preparation of protein will precipitate at a specific point rather than over a broad range of solution conditions. Furthermore, degradation during storage and/or crystallization is minimized if all of the proteases have been removed.

Although there are a number of ways to check the purity of a protein, the most convenient, and widely used, involve electrophoresis. Most experimentalists use SDS–PAGE and/or isoelectric focusing to determine the purity and homogeneity of the protein. SDS–PAGE may be slightly more convenient for the detection of unrelated proteins; isoelectric focusing is probably more useful in detecting subspecies of the recombinant protein of interest. We will consider the nature and origins of such subspecies below. Once the protein(s) is fractionated, either on an isoelectric focusing gel or on SDS–PAGE, it is detected by staining, either with silver or with Coomassie brilliant blue. Neither reagent reacts uniformly with all proteins; depending on the proteins involved, either method can overestimate or underestimate the level of a contaminant relative to the desired recombinant protein. Silver staining is the more sensitive method. However, if there is sufficient material for a serious attempt at crystallography, the sensitivity of Coomassie staining is usually more than sufficient for analytical purposes. It is often useful to fractionate a protein preparation by both isoelectric focusing and SDS–PAGE, and stain gels with silver and Coomassie brilliant blue. This increases the chance of discovery of an important contaminant and/or heterogeneity in the protein preparation.

If the preparation is relatively free of unrelated proteins, but there is concern about the presence of multiple species of the desired recombinant protein, there are several techniques that can be applied. Mass spectrometry is capable of detecting small differences in molecular weights, and for proteins up to several hundred amino acids in length it is usually able to detect differences in mass equivalent to a single amino acid. This can be useful in detecting heterogeneity in post-translational modifications, if such are present, and in detecting heterogeneity at both the amino and carboxyl termini. Amino-terminal sequencing can also be used to detect N-terminal heterogeneity, but has some limitations that are discussed below.

In *E. coli*, the methionine used to initiate translation is modified with a formyl group. The formyl group, and sometimes the amino-terminal methionine, is removed from proteins expressed in *E. coli*. Removal of the N-terminal amino acid is dependent on the identity of the second amino acid; methionines preceding small amino acids (Ala, Ser, Gly, Pro, Thr, Val) are generally removed (Waller, 1963; Tsunasawa *et al.*, 1985). However, when large amounts of a recombinant protein are made in *E. coli*, the formylase and aminopeptidase that mediate N-terminal processing are sometimes overwhelmed, and removal of the N-terminal groups is often incomplete. It is common to observe heterogeneity at the amino termini of even the most highly purified recombinant proteins. Amino-terminal sequencing can be used to detect this type of amino-terminal heterogeneity; however, the portion of the protein that retains the formyl group will not be detected by this method, and a misleading impression of the quantity and quality of the protein preparation can be obtained.

Heterogeneity at both the amino and carboxyl termini can be introduced by proteolysis, especially when the ends of the protein are extended and unstructured. This problem is frequently encountered when domains (rather than intact proteins) are expressed and can often be avoided if the boundaries of compact structural domains are precisely defined. In addition to introducing heterogeneity due to partial proteolysis, dangling ends can contribute to aggregation.

In terms of crystallization, the ability to produce a highly concentrated monodisperse protein preparation is probably more important than absolute purity. There are a number of techniques that can be used to determine whether or not the protein is aggregating. Analytical ultracentrifugation is the classical method, and size-exclusion chromatography has been widely used, particularly by biochemists. However, many crystallographers routinely use dynamic light scattering to check concentrated protein preparations for aggregation (Ferré-D'Amaré & Burley, 1994). The method is relatively simple, very sensitive to small amounts of aggregation and has the additional advantage that it does not consume the sample. After testing, the sample (which is often precious) can still be used for crystallization trials.

If sample heterogeneity is detected, one is faced with the issue of whether it will adversely affect crystallization, and if so, how to remove it. Unfortunately, there do not seem to be general rules. Heterogeneity at the termini of proteins is a common occurrence. In many crystal structures, the termini are disordered and heterogeneity at these unstructured ends would not be expected to be a significant problem. Indeed, in a number of instances, N-terminal sequence analysis of proteins obtained by dissolving crystals has indicated substantial heterogeneity. However, in other cases, properly defined domain boundaries are thought to have been a critical factor in obtaining useful crystals. Domain boundaries can be determined by a combination of limited proteolysis, followed by identification of the fragments using mass spectrometry (Cohen *et al.*, 1995; Hubbard, 1998). Subsequent re-engineering of expression constructs with modified termini is a relatively easy task. Similar engineering can also be used to alter internal sequences, such as removal of sites of post-translational modification or introduction of mutations that improve solubility (Chapter 4.3).

3.1.6.2. Protein storage

Even when the efforts of those engaged in crystallization and those engaged in producing the desired recombinant protein are well coordinated, it is not usually appropriate or desirable to use all the available protein for crystallization at the same time. This means that some of the material must be stored for later use. Even under the best of circumstances, protein solutions are subject to a number of unwanted events that can include, but are not limited to, oxidation, racemization, deamination, denaturation, proteolysis and aggregation. As a general rule, it is better to store proteins as highly purified concentrated solutions. This reduces problems of proteolysis (since the proteases have been removed), and, in general, proteins are better behaved if they are relatively concentrated (greater than 1 mg ml^{-1}). This is not an absolute rule, however; if there are problems with aggregation, these can sometimes be minimized by storage of proteins in dilute solutions, followed by concentration of the samples immediately prior to crystallization. If the protein contains oxidizable sulfurs (free cysteines are a particular problem), reducing agents can be added (and should be refreshed as necessary), and the solutions held in a non-reducing (N_2) atmosphere. In some cases, it is easier to mutate surface cysteines to produce a more stable protein (see Chapter 4.3).

In general, proteins behave best under conditions of pH and ionic strength similar to those they would experience in the normal host. Usually this means a pH near, or slightly above, neutral and intermediate ionic strength. These conditions are often not the ideal conditions for crystallization, and dialysis or other forms of buffer exchange may be required before beginning crystallization trials. In general, protein solutions are stored either at 4 °C in a cold room or refrigerator, or at 0 °C on ice. It is essential that the protein be stored in a manner that will not allow microbial growth, usually achieved by sterilization of the protein solution by filtration through 0.2 micron filters and/or addition of antimicrobial agents, such as NaN_3. For long-term storage (periods longer than a few weeks), protein solutions are often precipitated in ammonium sulfate or frozen at either −20 or −70 °C. Repeated freezing and thawing is not recommended; if a protein sample is to be frozen, it should be divided into aliquots small enough so that each will be thawed only once. Whenever a protein sample is frozen and thawed, some loss of quality and/or activity can be expected. Freezing samples of intermediate concentration (1–3 mg ml^{-1}) usually works better than freezing either extremely dilute or concentrated samples. Cryoprotective agents can be added to protein samples destined to be frozen; however, it should be remembered that the same reagents that are helpful when freezing a protein sample may be distinctly unhelpful when that sample is thawed and used for crystallography. Most biochemists willingly add glycerol to their protein samples before freezing; crystallographers are not usually happy to find that their protein sample is dissolved in 50% glycerol. Both pH and ionic strength can affect a protein's tolerance to freezing and thawing. In many cases, buffer exchange and concentration procedures need to be performed to convert stored protein solutions to ones suitable for crystallization.

As is so often true in science, decisions about whether to freeze a particular protein sample and, if it is to be frozen, exactly how the freezing should be done, depend on experience. If the protein in question is an enzyme, it is often useful to set up a series of trials in which small aliquots of the protein are stored under a variety of conditions. If the aliquots are tested on a fairly regular basis, how stable the protein is in solution can usually be determined, as well as how well it will tolerate a cycle of freezing and thawing, with or without an added cryoprotectant. If enzyme assays are not available, other methods of characterization, such as gel electrophoresis, mass spectrometry and light scattering, can be used to check for degradation, oxidation of cysteines and aggregation. Armed with this information, and with a plan for how the protein will be used for crystallization, it is usually a fairly simple matter to decide whether or not to freeze a particular sample, and, if the sample is to be frozen, how best to do it. It is a good idea to make such tests early in a major crystallization effort. This will avoid the awkward dilemma that occurs when a large amount of a highly purified protein is available, and the knowledge of how best to store it is not.

3.1.7. Reprise

We have reached a point where it is possible to use recombinant DNA techniques to produce most proteins in quantities sufficient for crystallography. Both high-level expression systems and methods for making defined modifications of recombinant proteins vastly simplify the process of purification. This has played a direct and critical role in the ability of crystallographers to produce an astonishing array of new and exciting protein structures. We are beginning to come to grips with the next level of the problem: using the ability to modify the sequence of proteins to improve their crystallization properties. This is a difficult problem; however, there are already notable, if hard won, successes. It would appear that the marriage of genetic engineering and crystallography – clearly a case in which opposites attract – has been a happy union. This is entirely for the good. Collaborations between specialists in these disciplines have led to the solution of problems too difficult for any individual armed only with the skills of one or the other partner. It is important that genetic engineering be fully integrated into future crystallographic efforts, either directly within the crystallography laboratory or through close collaborations. There yet remain formidable problems in protein structure and function that will require all the combined talents of the most skilled practitioners of these arcane arts.

References

Abelson, J. N. & Simon, M. I. (1990). *Guide to protein purification. Methods Enzymol.* **182**, 1–894.

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A. & Struhl, K. (1995). *Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology*, 3rd ed. New York: Greene Publishing Associates and Wiley.

Beggs, J. D. (1978). *Transformation of yeast by a replicating hybrid plasmid. Nature (London)*, **275**, 104–109.

Bhandari, P. & Gowrishankar, J. (1997). *An Escherichia coli host strain useful for efficient overproduction of cloned gene products with NaCl as the inducer. J. Bacteriol.* **179**, 4403–4406.

Biswas, E. E., Fricke, W. M., Chen, P. H. & Biswas, S. B. (1997). *Yeast DNA helicase A: cloning, expression, purification, and enzymatic characterization. Biochemistry*, **36**, 13277–13284.

Bollag, D. M., Rozycki, M. D. & Edelstein, S. J. (1996). *Protein Methods.* New York: Wiley-Liss.

Boyer, P. L. & Hughes, S. H. (1996). *Site-directed mutagenic analysis of viral polymerases and related proteins. Methods Enzymol.* **275**, 538–555.

Brinkmann, U., Mattes, R. E. & Buckel, P. (1989). *High-level expression of recombinant genes in Escherichia coli is dependent on the availability of the dnaY gene product. Gene*, **85**, 109–114.

Broach, J. R. (1983). *Construction of high copy number yeast vectors using 2 μm circle sequences. Methods Enzymol.* **101**, 307–325.

Chong, S., Mersha, F. B., Comb, D. G., Scott, M. E., Landry, D., Vence, L. M., Perler, F. B., Benner, J., Kucera, R. B., Hirvonen, C. A., Pelletier, J. J., Paulus, H. & Xu, M. Q. (1997). *Single-column purification of free recombinant proteins using a self-cleavable affinity tag derived from a protein splicing element. Gene*, **192**, 271–281.

Chong, S., Shao, Y., Paulus, H., Benner, J., Perler, F. B. & Xu, M. Q. (1996). *Protein splicing involving the Saccharomyces cerevisiae VMA intein. The steps in the splicing pathway, side reactions leading to protein cleavage, and establishment of an in vitro splicing system. J. Biol. Chem.* **271**, 22159–22168.

Cohen, S. L., Ferre-D'Amare, A. R., Burley, S. K. & Chait, B. T. (1995). *Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. Protein Sci.* **46**, 1088–1099.

Cole, P. A. (1996). *Chaperone-assisted protein expression. Structure*, **4**, 239–242.

Cregg, J. M., Vedick, T. S. & Raschke, W. C. (1993). *Recent advances in the expression of foreign genes in Pichia pastoris. Biotechnology*, **11**, 905–910.

De Bernardez Clark, E. (1998). *Refolding of recombinant proteins. Curr. Opin. Biotechnol.* **9**, 157–163.

De Boer, H. A. & Kastelein, R. A. (1986). *Biased codon usage: an exploration of its role in optimization of translation.* In *Maximizing Gene Expression*, edited by W. S. Reznikioff & L. Gold, pp. 225–285. Boston: Butterworths.

Del Tito, B. J. Jr, Ward, J. M., Hodgson, J., Gershater, C. J. L., Edwards, H., Wysocki, L. A., Watson, F. A., Sathe, G. & Kane, J. F. (1995). *Effects of a minor isoleucyl tRNA on heterologous protein translation in Escherichia coli. J. Bacteriol.* **177**, 7086–7091.

Enfors, S.-O. (1992). *Control of in vivo proteolysis in the production of recombinant proteins. Trends Biotechnol.* **10**, 310–315.

Ernst, J. F. & Kawashima, E. (1988). *Variations in codon usage are not correlated with heterologous gene expression in Saccharomyces cerevisiae and Escherichia coli. J. Biotechnol.* **7**, 1–9.

Ferré-D'Amaré, A. R. & Burley, S. K. (1994). *Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. Structure*, **2**, 357–359.

Fischer, B., Sumner, I. & Goodenough, P. (1993). *Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. Biotechnol. Bioeng.* **41**, 3–13.

Georgiou, G. & Valax, P. (1996). *Expression of correctly folded proteins in Escherichia coli. Curr. Opin. Biotechnol.* **7**, 190–197.

Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981). *Translational initiation in prokaryotes. Annu. Rev. Microbiol.* **35**, 365–403.

Goldman, E., Rosenberg, A. H., Zubay, G. & Studier, F. W. (1995). *Consecutive low-usage leucine codons block translation only when near the 5′ end of a message in Escherichia coli. J. Mol. Biol.* **245**, 467–473.

Gottesman, S. (1990). *Minimizing proteolysis in Escherichia coli: genetic solutions. Methods Enzymol.* **185**, 119–129.

Grinna, L. S. & Tschopp, J. F. (1989). *Size distribution and general structural features of N-linked oligosaccharides from the methylotrophic yeast, Pichia pastoris. Yeast*, **5**, 107–115.

Guise, A. D., West, S. M. & Chaudhuri, J. B. (1996). *Protein folding in vivo and renaturation of recombinant proteins from inclusion bodies. Mol. Biotechnol.* **6**, 53–64.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J.* **9**, 1665–1672.

Hernan, R. A., Hui, H. L., Andracki, M. E., Noble, R. W., Sligar, S. G., Walder, J. A. & Walder, R. Y. (1992). *Human hemoglobin expression in Escherichia coli: importance of optimal codon usage. Biochemistry*, **31**, 8619–8628.

Higgins, D. R. & Cregg, J. (1998). *Methods in Molecular Biology*, Vol. 103. *Pichia protocols.* Totowa: Humana Press.

Hirel, P. H., Schmitter, M. J., Dessen, P., Fayat, G. & Blanquet, S. (1989). *Extent of N-terminal methionine excision from Escherichia coli proteins is governed by the side-chain length of the penultimate amino acid. Proc. Natl Acad. Sci. USA*, **86**, 8247–8251.

Hockney, R. C. (1994). *Recent developments in heterologous protein production in Escherichia coli. Trends Biotechnol.* **12**, 456–463.

Hofmann, A., Tai, M., Wong, W. & Glabe, C. G. (1995). *A sparse matrix screen to establish initial conditions for protein renaturation. Anal. Biochem.* **230**, 8–15.

Hollenberg, C. P. & Gellissen, G. (1997). *Production of recombinant proteins by methylotrophic yeasts. Curr. Opin. Biotechnol.* **8**, 554–560.

Hubbard, S. J. (1998). *The structural aspects of limited proteolysis of native proteins. Biochim. Biophys. Acta*, **1382**, 191–206.

Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (1990). *PCR Protocols: A Guide to Methods and Applications.* San Diego: Academic Press.

Jarvis, D. L., Kawar, Z. S. & Hollister, J. R. (1998). *Engineering N-glycosylation pathways in the baculovirus-insect cell system. Curr. Opin. Biotechnol.* **9**, 528–533.

Jones, I. & Morikawa, Y. (1996). *Baculovirus vectors for expression in insect cells. Curr. Opin. Biotechnol.* **7**, 512–516.

Kane, J. F. (1995). *Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr. Opin. Biotechnol.* **6**, 494–500.

Kaufman, R. J. (1990). *Selection and coamplification of heterologous genes in mammalian cells. Methods Enzymol.* **185**, 537–566.

Kim, R., Sandler, S. J., Goldman, S., Yokota, H., Clark, A. J. & Kim, S.-H. (1998). *Overexpression of archaeal proteins in Escherichia coli. Biotechnol. Lett.* **20**, 207–210.

Krueger, J. K., Kulke, M. H., Schutt, C. & Stock, J. (1989). *Protein inclusion body formation and purification. BioPharm*, March issue, 40–45.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature (London)*, **393**, 648–659.

LaVallie, E. R., DiBlasio, E. A., Kovacic, S., Grant, K. L., Schendel, P. F. & McCoy, J. M. (1993). *A thioredoxin gene fusion expression system that circumvents inclusion body formation in the E. coli cytoplasm. Biotechnology*, **11**, 187–193.

LaVallie, E. R. & McCoy, J. M. (1995). *Gene fusion expression systems in Esherichia coli. Curr. Opin. Biotechnol.* **6**, 501–506.

Lee, H. W., Joo, J.-H., Kang, S., Song, L.-S., Kwon, J.-B., Han, M. H. & Na, D. S. (1992). *Expression of human interleukin-2 from native and synthetic genes in E. coli: no correlation between major codon bias and high level expression. Biotechnol. Lett.* **14**, 653–658.

Lu, A. & Miller, L. K. (1996). *Generation of recombinant baculoviruses by direct cloning. Biotechniques*, **21**, 63–68.

McCarroll, L. & King, L. A. (1997). *Stable insect cell cultures for recombinant protein production. Curr. Opin. Biotechnol.* **8**, 590–594.

McPherson, M. J., Hames, B. D. & Taylor, G. R. (1995). *PCR 2: A Practical Approach*. Oxford, New York: IRL Press at Oxford University Press.

Makrides, S. C. (1996). *Strategies for achieving high-level expression of genes in Escherichia coli. Microbiol. Rev.* **60**, 512–538.

Marston, F. A. (1986). *The purification of eukaryotic polypeptides synthesized in Escherichia coli. Biochem. J.* **240**, 1–12.

Merrington, C. L., Bailey, M. J. & Possee, R. D. (1997). *Manipulation of baculovirus vectors. Mol. Biotechnol.* **8**, 283–297.

Mitraki, A. & King, J. (1989). *Protein folding intermediates and inclusion body formation. Biotechnology*, **7**, 690–697.

Mohsen, A.-W. A. & Vockley, J. (1995). *High-level expression of an altered cDNA encoding human isovaleryl-CoA dehydrogenase in Escherichia coli. Gene*, **160**, 263–267.

Murby, M., Uhlén, M. & Ståhl, S. (1996). *Upstream strategies to minimize proteolytic degradation upon recombinant production in Escherichia coli. Protein Expr. Purif.* **7**, 129–136.

Nilsson, B., Forsberg, G., Moks, T., Hartmanis, M. & Uhlén, M. (1992). *Fusion proteins in biotechnology and structural biology. Curr. Opin. Struct. Biol.* **2**, 569–575.

O'Reilly, D. R., Miller, L. K. & Luckow, V. A. (1992). *Baculovirus Expression Vectors: A Laboratory Manual.* New York: W. H. Freeman & Co.

Pfeifer, T. A. (1998). *Expression of heterologous proteins in stable insect culture. Curr. Opin. Biotechnol.* **9**, 518–521.

Possee, R. D. (1997). *Baculoviruses as expression vectors. Curr. Opin. Biotechnol.* **7**, 569–572.

Richardson, C. D. (1995). *Methods in Molecular Biology*, Vol. 39. *Baculovirus Expression Protocols*. Totowa: Humana Press.

Richarme, G. & Caldas, T. D. (1997). *Chaperone properties of the bacterial periplasmic substrate-binding proteins. J. Biol. Chem.* **272**, 15607–15612.

Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. D. & Gold, L. (1992). *Translational initiation in Escherichia coli: sequences within the ribosome-binding site. Mol. Microbiol.* **6**, 1219–1229.

Romanos, M. (1995). *Advances in the use of Pichia pastoris for high-level gene expression. Curr. Opin. Biotechnol.* **6**, 527–533.

Romanos, M. A., Scorer, C. A. & Clare, J. J. (1992). *Foreign gene expression in yeast: a review. Yeast*, **8**, 423–488.

Rossi, F. M. & Blau, H. M. (1998). *Recent advances in inducible expression systems. Curr. Opin. Biotechnol.* **9**, 451–456.

Sachdev, D. & Chirgwin, J. M. (1998). *Solubility of proteins isolated from inclusion bodies is enhanced by fusion to maltose-binding protein or thioredoxin. Protein Expr. Purif.* **12**, 122–132.

Saez, E., No, D., West, A. & Evans, R. M. (1997). *Inducible expression in mammalian cells and transgenic mice. Curr. Opin. Biotechnol.* **8**, 608–616.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd ed. New York: Cold Spring Harbor Laboratory Press.

Samuelsson, E., Moks, T., Nilsson, B. & Uhlén, M. (1994). *Enhanced in vitro refolding of insulin-like growth factor I using a solubilizing fusion partner. Biochemistry*, **33**, 4207–4211.

Schein, C. H. & Noteborn, M. H. M. (1988). *Formation of soluble recombinant proteins in Escherichia coli is favored by lower growth temperature. Biotechnology*, **6**, 291–294.

Schenk, P. M., Baumann, S., Mattes, R. & Steinbiss, H.-H. (1995). *Improved high-level expression system for eukaryotic genes in Escherichia coli using T7 RNA polymerase and rare ArgtRNAs. Biotechniques*, **19**, 196–198.

Sclimenti, C. R. & Calos, M. P. (1998). *Epstein–Barr virus vectors for gene expression and transfer. Curr. Opin. Biotechnol.* **9**, 476–479.

Scopes, R. K. (1994). *Protein Purification: Principles and Practice*. New York: Springer-Verlag.

Shimotohno, K. & Temin, H. M. (1982). *Loss of intervening sequences in mouse α-globin DNA inserted in an infectious retrovirus vector. Nature (London)*, **299**, 265–268.

Sorge, J. & Hughes, S. H. (1982). *Splicing of intervening sequences introduced into an infectious retroviral vector. J. Mol. Appl. Genet.* **1**, 547–559.

Studier, F. W. & Moffatt, B. A. (1986). *Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol.* **189**, 113–130.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol.* **185**, 60–89.

Tabor, S. & Richardson, C. C. (1985). *A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. Proc. Natl Acad. Sci. USA*, **82**, 1074–1078.

Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. (1991). *The N-end rule in bacteria. Science*, **254**, 1374–1377.

Tsunasawa, S., Stewart, J. W. & Sherman, F. S. (1985). *Amino-terminal processing of mutant forms of yeast iso-1-cytochrome c. The specificities of methionine aminopeptidase and acetyltransferase. J. Biol. Chem.* **260**, 5382–5391.

Unger, T. F. (1997). *Show me the money: prokaryotic expression vectors and purification systems. The Scientist*, **11**, 20–23.

Wall, J. G. & Plückthun, A. (1995). *Effects of overexpressing folding modulators on the in vivo folding of heterologous proteins in Escherichia coli. Curr. Opin. Biotechnol.* **6**, 507–516.

Waller, J.-P. (1963). *The NH_2-terminal residues of the proteins from cell-free extracts of E. coli. J. Mol. Biol.* **7**, 483–496.

Werner, M. H., Clore, G. M., Gronenborn, A. M., Kondoh, A. & Fisher, R. J. (1994). *Refolding proteins by gel filtration chromatography. FEBS Lett.* **345**, 125–130.

Wilkinson, D. L., Ma, N. T., Haught, C. & Harrison, R. G. A. (1995). *Purification by immobilized metal affinity chromatography of human atrial natriuretic peptide expressed in a novel thioredoxin fusion protein. Biotechnol. Prog.* **11**, 265–269.

Yasukawa, T., Kanei-Ishii, C., Maekawa, T., Fujimoto, J., Yamamoto, T. & Ishii, S. (1995). *Increase of solubility of foreign proteins in Escherichia coli by coproduction of the bacterial thioredoxin. J. Biol. Chem.* **270**, 25328–25331.

Yonemoto, W. M., McGlone, M. L., Slice, L. W. & Taylor, S. S. (1998). *Prokaryotic expression of catalytic subunit of adenosine cyclic monophosphate-dependent protein kinase.* In *Protein Phosphorylation*, edited by B. M. Sefton & T. Hunter, pp. 419–434. San Diego: Academic Press.

Zhang, S. P., Zubay, G. & Goldman, E. (1991). *Low usage codons in Escherichia coli, yeast, fruit fly and primates. Gene*, **105**, 61–72.

International Tables for Crystallography (2012). Vol. F, Chapter 3.2, pp. 92–98.

Chapter 3.2. Expression and purification of membrane proteins for structural studies

J. A. Ernst, D. G. Yansura and C. M. Koth

3.2.1. Introduction

Integral membrane proteins constitute about a third of the proteome of most organisms but less than 1% of all entries in protein structural databases (Berman *et al.*, 2002). This disparity is largely due to inherent difficulties in their expression, solubilization and purification. The production of sufficient protein for structural studies can be challenging for any target, but several obstacles are unique to membrane proteins. For example, proper insertion into the membrane relies on host cellular machinery that may be limiting or incompatible. Exceeding this capacity can lead to cell death or the accumulation of aggregated and inactive protein within the cell (Geertsma *et al.*, 2008). Also, post-translational modifications such as glycosylation, acylation and sulfation may not be faithfully reproduced (Grisshammer & Tate, 1995). As a further complication, even if sufficient expression can be achieved, most membrane-protein structural studies require that the target be extracted from the cellular membrane using detergents. However, detergents can adversely affect protein structure and function, as well as influence the outcome of crystal trials (Engel *et al.*, 2002; Lemieux *et al.*, 2003; Prive, 2007). Fortunately, if a strategy for purifying sufficient quantities of a given membrane-protein target *can* be established, crystallization strategies largely mimic the standard techniques for soluble proteins (Newby *et al.*, 2009).

Despite the aforementioned challenges, high-resolution structures of almost 200 unique membrane proteins have now been solved (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html), the vast majority using protein produced by recombinant methods (Willis & Koth, 2008). The rate of new structure determinations has also increased dramatically over the last few years, mimicking the exponential growth of soluble protein structures in the early 1980s. Space limitations here preclude a thorough review of all possible membrane-protein production methods. However, an examination of the successful expression, solubilization and purification strategies that have led to membrane-protein structures reveals that, in many cases, remarkably similar methods have been used (Carpenter *et al.*, 2008; Willis & Koth, 2008; Newby *et al.*, 2009). Guided by these methods, the following sections detail a rational and consensus 'first-attempt' strategy that has worked for a broad range of membrane targets with only minor variations in technique (Dobrovetsky *et al.*, 2005, 2007; Lunin *et al.*, 2006). It must be noted that while this represents an evidence-guided approach, the methods provided herein will not work for every membrane protein and, in fact, will fail for many. Thus, prioritized lists of alternative strategies are provided for those targets in which the initial expression or isolation attempts do not succeed or problems are encountered.

3.2.2. A consensus strategy for membrane-protein expression

Historically, the likelihood of successful structure determination has been high for those membrane proteins that can be isolated from readily available abundant natural sources (Sakai & Tsukihara, 1998). Unfortunately, most membrane proteins do not meet this criterion. In reality, the vast majority of high-resolution membrane-protein structures are of prokaryotic targets, expressed in *Escherichia coli* by recombinant methods (Willis & Koth, 2008; Newby *et al.*, 2009). The reasons for this are simple. Attempting recombinant expression in *E. coli* is inexpensive, flexible, simple and easily scaled-up, and many constructs and strains can be screened quickly. As with crystallographic efforts for many soluble proteins (Gräslund *et al.*, 2008), most successful membrane-protein endeavours also commonly screen multiple constructs and/or orthologues for any given target (see, for example, Chang *et al.*, 1998; Doyle *et al.*, 1998), since the greater the number of unique constructs screened, the greater the chance that one will be successfully isolated and crystallized. Given the overwhelming use of *E. coli* for successful membrane-protein structures, a 'first-pass' expression strategy suitable for almost any membrane protein is clear: attempt expression in *E. coli* and, whenever possible, screen multiple constructs, orthologues and strains.

3.2.2.1. *Choosing the expression system and affinity tags*

In addition to the choice of expression host, one must also consider the expression system and the type and placement of affinity tags or fusion proteins. For *E. coli*, IPTG-inducible T7 polymerase-driven expression systems, such as those based on pET vectors (Studier & Moffatt, 1986) and λDE3 lysogen strains, are the most widely used for membrane proteins, as is the case for soluble targets. Also, it has generally been observed that for most targets, protein expression is optimal at lower temperatures (*i.e.* <20 °C; Christendat *et al.*, 2000; Wang *et al.*, 2003; Dobrovetsky *et al.*, 2005). When structural studies are the desired outcome, the most common tagging strategy for membrane proteins is to engineer a stretch of at least six histidine residues at the amino or carboxyl terminus of the target constructs; this is used for >80% of successful targets (Willis & Koth, 2008). If expression levels are sufficient (*i.e.* ≥ 0.05 mg g^{-1} cell paste), this often permits purification using a general two-step procedure: capture of the tagged protein by immobilized metal affinity chromatography (IMAC), followed by size-exclusion chromatography (SEC) (see below). Remarkably, this basic approach has proved successful for the crystallization of many membrane proteins, as discussed in the following section. If premature termination is observed during expression, the engineering of a carboxyl-terminal tag will ensure that these proteins are not isolated during purification. Also, extending the stretch of histidines to greater than six residues can improve the retention of membrane proteins on immobilized metal affinity resins. This can prove particularly useful, given the modest expression levels and reduced chromatographic resolution and recovery of many membrane targets (Dobrovetsky *et al.*, 2005; Eshaghi *et al.*, 2005; Surade *et al.*, 2006; Lewinson *et al.*, 2008). Other affinity tags are viable options, but are more rarely used. For example, although fusion proteins such as glutathione S-transferase (GST) and thioredoxin are widely used to promote expression and/or simplify the purification of soluble proteins [for an extensive comparison of various affinity tags, see Lichty *et al.* (2005)], they are rarely cited in expression strategies for polytopic membrane proteins.

Copyright © 2012 International Union of Crystallography

Table 3.2.2.1

Strategies for improving recombinant membrane-protein expression in *E. coli*

Troubleshooting expression problems: if the 'first-pass' expression strategy is unsuccessful, then some changes in the expression construct, host strain or induction conditions may be helpful. The symptoms of the problem often give a clue as to how to make these corrections.

Symptom	Solution
E. coli colonies small after transformation into expression strain	Need tighter promoter such as phoA, araC *etc.*
Unprocessed signal sequence†	If none is used in native situation, remove from construct† If one is present in native situation, use an expression host co-translational signal sequence† If still unprocessed, lower translation initiation strength
Mostly high-molecular-weight aggregates	Decrease induction time
Rapid cell growth after induction with low expression levels	Increase translation initiation strength
Poor cell growth after induction with low expression levels	
(*a*) Accumulation maximum after only 1–3 h	Decrease translation initiation rates or plasmid copy number
(*b*) Accumulation increases out to 8–10 h post induction	Usually successful: (*a*) Re-synthesize the gene, removing rare codons and optimizing codon pairs† (*b*) Convert transmembrane and intracellular domain cysteines (normally reduced) to serines† (*c*) Try different non-DE3 lysogen strains after switching to a non-T7 promoter (i) Cytoplasmic protease deletions (ii) Periplasmic protease deletions (iii) Heat-shock htpR^{TS} Sometimes successful: (*a*) Try different C-terminal tags such as poly-his, Flag *etc.*† (*b*) Try different temperatures of induction such as 20, 25, 30 °C

† These methods are also useful for optimizing eukaryotic gene expression.

3.2.2.2. Strategies for improving membrane-protein expression

Low expression levels are common for many membrane proteins. Before testing alternative hosts, it may be worthwhile first to screen alternative *E. coli* expression systems, such as those using the tightly controlled arabinose promoter (Guzman *et al.*, 1995). Also, other *E. coli* strains can be tested. For example, the mutant C41 and C43 'Walker' expression strains (Miroux & Walker, 1996; Wagner *et al.*, 2008) yield higher expression levels for some membrane proteins. Table 3.2.2.1 summarizes common strategies for improving the expression of membrane proteins in *E. coli*. If reasonable efforts in *E. coli* still do not yield sufficient expression, then an alternative expression host should be considered. The next most successful is yeast. Of the 22 recombinant eukaryotic membrane proteins whose structures have been solved by crystallography, half were produced using *Pichia pastoris* (seven targets), *Saccharomyces cerevisiae* (three targets) or *Saccharomyces pombe* (one target) (Raman *et al.*, 2006). Membrane proteins expressed in yeast are typically fused to an amino-terminal host signal sequence to promote proper membrane targeting. The use of higher eukaryotic hosts, such as insect and mammalian cells, is also increasing as more groups attempt to crystallize recombinant eukaryotic targets. The recent high-resolution structures of several G-protein-coupled receptors are notable examples of targets expressed successfully in insect cells (Cherezov *et al.*, 2007; Rasmussen *et al.*, 2007; Hanson *et al.*, 2008; Jaakola *et al.*, 2008; Warne *et al.*, 2008).

To date, there is only one report of a structure determined from cell-free synthesis of the target membrane protein (EmrE; Chen *et al.*, 2007). However, for several reasons, cell-free approaches should be explored as a complement to traditional expression methods. Firstly, the absence of a host cell negates toxicity problems for some target proteins. Secondly, many targets that do not express in cells can be produced cell-free (Ishihara *et al.*, 2005; Klammt *et al.*, 2007; Savage *et al.*, 2007). Thirdly, these systems allow for considerable flexibility with respect to reaction conditions and components. Cell-free systems for prokaryotic and eukaryotic hosts have been described (Schwarz *et al.*, 2008), some of which function in the presence of solubilizing detergents (Ishihara *et al.*, 2005; Klammt *et al.*, 2007; Savage *et al.*, 2007) or lipids (Kuruma *et al.*, 2005; Schwarz *et al.*, 2007), facilitating proper folding and downstream purification of membrane-protein targets.

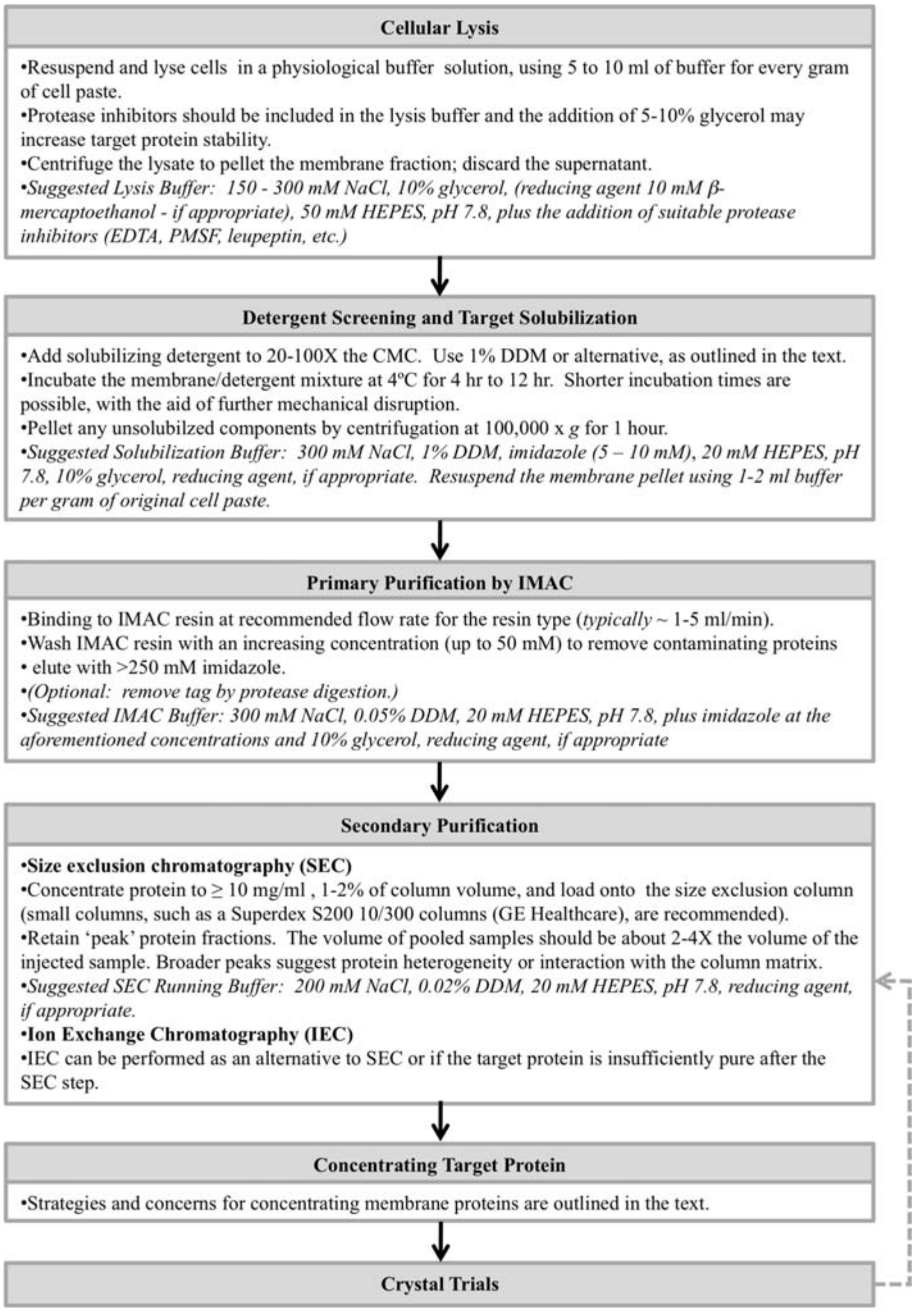
Many eukaryotic proteins have undergone extensive construct engineering to improve expression and/or crystallization outcomes. These include fusion to T4 lysozyme, mutation of glycosylation sites, deletion of potentially disordered regions and/or targeted evolution to identify conformationally stable mutants (Long *et al.*, 2005; Cherezov *et al.*, 2007; Rosenbaum *et al.*, 2007; Jaakola *et al.*, 2008; Magnani *et al.*, 2008; Serrano-Vega *et al.*, 2008; Warne *et al.*, 2009). Also worth mentioning is the observation that incorporation of an amino-terminal 'rhodopsin tag' comprising the first 20 amino acids of bovine rhodopsin (amino-acid sequence MNGTEGPNFYVPFSNKTGVV) has been found to boost dramatically the expression levels of eukaryotic membrane targets expressed in insect and mammalian cells (Krautwurst *et al.*, 1998).

3.2.3. A consensus strategy for membrane-protein purification

3.2.3.1. General principles

The purification of an expressed integral membrane protein for structural studies typically involves four discrete steps: cell lysis, membrane isolation, detergent extraction and chromatographic separation. In this section, an evidence-based consensus strategy is outlined which aims to serve as a starting point for membrane-protein purification. Detailed protocols are outlined in Fig.

3.2.3.1. Note that the most significant complication in this process is the need to use detergents in all steps of protein handling and chromatography. Detergents permit the extraction of target proteins from the membrane, but they must also maintain protein stability and activity by mimicking, as closely as possible, the lipid bilayer. Although many detergents do meet these criteria, relatively few are compatible with high-resolution structural studies.



Cellular Lysis

- •Resuspend and lyse cells in a physiological buffer solution, using 5 to 10 ml of buffer for every gram of cell paste.
- •Protease inhibitors should be included in the lysis buffer and the addition of 5-10% glycerol may increase target protein stability.
- •Centrifuge the lysate to pellet the membrane fraction; discard the supernatant.
- •*Suggested Lysis Buffer: 150 - 300 mM NaCl, 10% glycerol, (reducing agent 10 mM β-mercaptoethanol - if appropriate), 50 mM HEPES, pH 7.8, plus the addition of suitable protease inhibitors (EDTA, PMSF, leupeptin, etc.)*

Detergent Screening and Target Solubilization

- •Add solubilizing detergent to 20-100X the CMC. Use 1% DDM or alternative, as outlined in the text.
- •Incubate the membrane/detergent mixture at 4ºC for 4 hr to 12 hr. Shorter incubation times are possible, with the aid of further mechanical disruption.
- •Pellet any unsolubilized components by centrifugation at 100,000 x *g* for 1 hour.
- •*Suggested Solubilization Buffer: 300 mM NaCl, 1% DDM, imidazole (5 – 10 mM), 20 mM HEPES, pH 7.8, 10% glycerol, reducing agent, if appropriate. Resuspend the membrane pellet using 1-2 ml buffer per gram of original cell paste.*

Primary Purification by IMAC

- •Binding to IMAC resin at recommended flow rate for the resin type (*typically* ~ 1-5 ml/min).
- •Wash IMAC resin with an increasing concentration (up to 50 mM) to remove contaminating proteins
- • elute with >250 mM imidazole.
- •*(Optional: remove tag by protease digestion.)*
- •*Suggested IMAC Buffer: 300 mM NaCl, 0.05% DDM, 20 mM HEPES, pH 7.8, plus imidazole at the aforementioned concentrations and 10% glycerol, reducing agent, if appropriate*

Secondary Purification

- •**Size exclusion chromatography (SEC)**
- •Concentrate protein to ≥ 10 mg/ml , 1-2% of column volume, and load onto the size exclusion column (small columns, such as a Superdex S200 10/300 columns (GE Healthcare), are recommended).
- •Retain 'peak' protein fractions. The volume of pooled samples should be about 2-4X the volume of the injected sample. Broader peaks suggest protein heterogeneity or interaction with the column matrix.
- •*Suggested SEC Running Buffer: 200 mM NaCl, 0.02% DDM, 20 mM HEPES, pH 7.8, reducing agent, if appropriate.*
- •**Ion Exchange Chromatography (IEC)**
- •IEC can be performed as an alternative to SEC or if the target protein is insufficiently pure after the SEC step.

Concentrating Target Protein

- •Strategies and concerns for concentrating membrane proteins are outlined in the text.

Crystal Trials

Figure 3.2.3.1

Flow diagram of the membrane-protein purification strategy. Typically, only two chromatographic steps (affinity chromatography and SEC) are required for proteins that are expressed at reasonably high levels with a hexa-histidine tag. Ion-exchange chromatography (IEC) methods are sometimes employed, although they are not as common as SEC. If no crystals are obtained in initial sparse matrix screens, the protein can be exchanged to a different detergent and crystal trials attempted again (dotted arrow). Suggested 'first-try' buffers are indicated at each step.

3.2.3.2. Cell lysis and membrane isolation

Cellular lysis methods for the isolation of integral membrane proteins are similar to those for soluble targets. Cells over-expressing the protein of interest are re-suspended in a lysis buffer lacking detergent. Typically, high-flow/high-pressure cell-disruption devices are used, such as a continuous-flow homogenizer (EmulsiFlex, Avestin, Canada) or microfluidizer (Microfluidics Inc., Newton, Massachusetts, USA). Although these two devices achieve lysis by different mechanisms, both can disrupt the cells of almost all commonly used expression hosts, including yeast, and permit the rapid processing of large volumes. Other cell-disruption methods, such as sonication or nitrogen cavitation, can also be used for eukaryotic cells and, with slightly lower efficiency, *E. coli*. Protease inhibitors are included at this stage, although they are often not necessary in the later stages of purification. Following lysis, the membrane fraction of the cells is pelleted by ultracentrifugation.

3.2.3.3. Detergent extraction of membrane proteins

Following centrifugation, the crude-membrane pellet is re-suspended in a native lysis buffer. Detergent is then added to extract the membrane proteins from the lipid bilayer, at a molarity well above the critical micelle concentration (CMC, the concentration above which detergent molecules aggregate to form micelles). It is also more common than not that the same detergent used for solubilization is also used for subsequent purification and crystallization (Willis & Koth, 2008), albeit at a much lower concentration (typically just above the CMC). Broadly speaking, detergents fall into three main categories: charged (ionic), zwitterionic and non-charged (polar). Detergents containing no net charge are more likely to solubilize membrane proteins in their native state and are therefore more widely used for structural studies. In addition, detergents with sugar residues as head groups have proved particularly successful for the crystallization of membrane proteins, as have polyoxyethylene monoalkylethers (C_nE_m; Prive, 2007; Newstead *et al.*, 2008). Of all the membrane-protein structures solved to date, almost 70% have used one of just six different detergents for solubilization (Newstead *et al.*, 2008; Willis & Koth, 2008); with the exception of dodecyl-*N*,*N*-dimethylamine *N*-oxide (LDAO), all are non-charged. These six are, starting with the most prevalent, dodecyl-*β*-D-maltopyranoside (DDM), LDAO, octyl-polyoxyethylene (C8POE), decyl-*β*-D-maltoside (DM), octyl-*β*-D-glucoside (OG) and Triton X-100. Of these, DDM has been used for one out of almost every four crystallized targets and should be tested with any new membrane protein. The general properties of detergents and their use for membrane-protein solubilization have been reviewed by Hjelmeland (1990) and Neugebauer (1990).

The ability of detergents to solubilize target proteins can be tested on a small scale. The concentration of detergent that is optimal for solubilization should be determined empirically, although this can quickly become a multidimensional problem. A simple starting strategy, based on the purification methods for successfully crystallized membrane proteins, is to screen the aforementioned detergents at 1% (DDM, LDAO, Triton X-100) or 2% (C8POE, DM, OG). Detergents that are effective at these concentrations can be further screened to determine the optimal solubilization conditions (*i.e.* by varying detergent concentration *versus* solubilization time). Here, 'optimal' refers to the conditions yielding the highest level of soluble active target protein, or, if no activity assay exists, simply the conditions yielding the most target protein (but see caveats below). Although the level of solubilized target protein usually increases with time, many membrane proteins show reduced stability in detergent. It is advantageous, and more common, to minimize solubilization times (*i.e.* <4 h), thereby facilitating a more rapid purification. Physical methods such as sonication and passage through a cell disruptor may also be used to speed the solubilization process. Target proteins are considered soluble if, in the presence of detergent, they remain in solution after being subjected to 100 000 *g* for 1 h.

3.2.3.4. *General considerations for monitoring membrane-protein activity*

Although often overlooked (surprisingly so), developing a means of following membrane-protein activity can prove critical when optimizing solubilization and purification conditions. Unfortunately, there is no universal method for assaying the activity of all solubilized or purified membrane proteins. For receptors, obvious assays include those based on monitoring ligand binding. For many other targets, such as those critically dependent on the vectorial nature of the membrane for activity (*i.e.* ion channels), there may be no direct assay in the detergent-solubilized state. On the premise that *any* assay is better than none, some simple biophysical analyses are worth mentioning. Firstly, some membrane targets that function as multimers have been shown to retain quaternary structure in sodium dodecyl sulfate (SDS) and during polyacrylamide gel electrophoresis (PAGE) (Cortes & Perozo, 1997; Prive, 2007). This observation can provide a convenient means of monitoring quaternary structure under various solubilization or purification conditions. It has also been suggested that proteins demonstrating this behaviour may, in general, be stable in a variety of detergents, as is the case for the KcsA potassium channel (Cortes & Perozo, 1997). Secondly, for those multimeric targets whose quaternary structure is not maintained in SDS, monitoring the degree of cross-linking in various detergents/conditions can provide for a simple, albeit indirect, structural probe (Sukharev *et al.*, 1999). Thirdly, monitoring the sensitivity of a target protein to limiting amounts of protease can also provide a simple means of probing 'activity' or structure under various solubilization/purification conditions. For example, binding to magnesium alters the protease sensitivity of detergent-solubilized CorA channels, and this can be easily monitored by SDS–PAGE analysis of digested samples (Payandeh & Pai, 2006). In theory, such an assay could be used to probe structural changes upon ligand binding for almost any target, so long as a ligand is known. The growing use of Fab antibody fragments as aids to membrane-protein crystallization provides a fourth assay for some targets (Hunte & Michel, 2002; Day *et al.*, 2007). Antibodies (or antibody fragments) that bind a target protein in a cellular environment should also be able to bind that protein in detergent, as evaluated by an assay such as immunoprecipitation. Ideally, such antibodies are western-negative (*i.e.* they recognize nonlinear peptide epitopes that would only be present in a properly folded protein). Lastly, membrane proteins containing disulfide bonds (common in extracellular domains) often migrate more rapidly in SDS–PAGE than the same protein under reducing conditions. This can be used as an assay for proper protein folding in the cell, although this method does not provide information about target-protein stability in detergent.

3.2.3.5. *Primary purification: affinity chromatography*

Given that most membrane proteins are cloned with a histidine tag, the most common method for their primary purification is immobilized metal affinity chromatography (IMAC), used in >80% of recombinant membrane-protein structures (Willis & Koth, 2008; Newby *et al.*, 2009). IMAC methods for membrane proteins do not differ significantly from those commonly used for soluble proteins, with the notable exception that detergent is required in all buffers. During the wash and elution steps, the detergent concentration is typically reduced to a level above the CMC that is only just sufficient to maintain target-protein solubility. Ideally, the ratio of micelles to membrane-protein molecules is $\sim$2 at this and subsequent purification stages [for an in-depth analysis, the reader is strongly encouraged to read Helenius *et al.* (1979) and Wiener (2004)]. The histidine tag may be removed after IMAC elution, although it is also worthwhile to proceed with a sample that retains the tag, since they provide critical lattice contacts in some protein crystals (Carson *et al.*, 2007).

3.2.3.6. *Secondary purification: size-exclusion and ion-exchange chromatography*

For most successful structures, target membrane proteins are sufficiently pure for crystal trials after a two-step purification: IMAC followed by size-exclusion chromatography (SEC) or ion-exchange chromatography (IEC) (Willis & Koth, 2008; Newby *et al.*, 2009). A detailed strategy is outlined in Fig. 3.2.3.1. SEC is the most common secondary (*i.e.* post-IMAC) method, represented in about half of all successful membrane-protein structures (Willis & Koth, 2008). IEC methods have been used less frequently, being employed for about one out of every four structures (Willis & Koth, 2008). The simple fact that some integral membrane proteins lack sufficient polar area to interact effectively with ion-exchange resins may account for the difference. Both methods permit the exchange of one detergent (*i.e.* that used for solubilization) for another or to a higher grade (a consideration, given the high cost of many detergents). Ion-exchange methods are likely to give a more complete detergent exchange, although this does not necessarily improve crystallization outcomes (Lemieux *et al.*, 2003). In fact, incomplete detergent exchange and/or incomplete removal of specifically bound lipids can prove beneficial in downstream crystal trials (Long *et al.*, 2007).

3.2.3.7. *Concentrating membrane-protein samples for crystallography*

Concentration of the target membrane protein is typically required prior to crystallization. Many different concentration methods can be used (*i.e.* those employed for soluble proteins), although some caution is warranted. Detergent will normally concentrate with the protein, since most detergent micelles are too large to pass through typical dialysis or concentration membranes. This can be detrimental to crystallization; high detergent concentrations can denature protein or limit diffraction (Wiener, 2004). As a general rule, the highest-molecular-weight cutoff membrane that does not permit passage of the target protein should be used, as this will allow the passage of at least some protein-free detergent micelles. Also, any size-exclusion steps should be performed at a relatively high protein concentration (10–20 mg ml^{-1} or higher) to minimize volumes. Purified protein may also be dialysed to exchange non-detergent buffer components, so long as the detergent concentration of the bulk solution is maintained just above the CMC. However, the dialysis of many detergents will proceed slowly (days to weeks). Accordingly, dialysis does not appear to be a widely used method for controlling detergent concentration prior to crystallization (reviewed by Willis & Koth, 2008).

3.2.4. Common purification pitfalls and prioritized alternative strategies

3.2.4.1. Poor solubility of the target protein

If the target protein does not remain soluble during purification, an obvious alternative strategy is to test different detergents or detergent combinations. Given that just a handful of different detergents have been used to solubilize and crystallize most membrane proteins, testing alternatives is usually not a daunting task. The most common method of detergent exchange is SEC (Fig. 3.2.3.1).

The zwitterionic lyso-lipid mimetic detergents such as FOS-CHOLINE 12 (dodecylphosphatidylcholine) or 14 (FC12 and FC14; Anatrace, Maumee, Ohio, USA) will solubilize many membrane proteins (Eshaghi *et al.*, 2005) and there are a few reports of their use in crystallography [MscS mechanosensitive channel (Bass *et al.*, 2002); protein-conducting channel (van den Berg *et al.*, 2004)]. Should solubilization screening with the more commonly used detergents prove ineffective, it may be possible to solubilize first with a 'stronger' more lipid-like detergent, such as FC12, and then exchange to a non-charged detergent, such as DDM, during purification.

3.2.4.2. The isolated target protein does not crystallize

Although an in-depth analysis of membrane-protein crystallization methods is beyond the scope of this chapter (see instead Chapter 4.2), there are many purification parameters that can be varied to improve crystallization outcomes should initial trials fail to yield 'hits' or well diffracting crystals. Perhaps the foremost method is to exchange the protein into an alternative detergent or detergent mixture, typically by SEC, and repeat the crystal trials. Since the vast majority of membrane-protein structures have used one of only a handful of detergents (see above), it is likely that sampling even just a small number of other detergents would dramatically increase the likelihood of obtaining or optimizing crystals. Also, an increase in the alkyl-chain length by one methylene group (*i.e.* nonyl- *versus* octyl-β-D-glucoside) often leads to an increase in protein stability, and this too can improve crystallization outcomes (Wiener, 2004). If significant covalent disulfide aggregation of the protein is observed during purification, alkylation of free cysteines should be considered during the early stages of target-protein isolation. Other parameters that could be varied include the solution pH, salt concentration and temperature, or supplementing with known ligands, to name just a few.

3.2.5. Summary

There are many possible routes for the expression and isolation of membrane proteins for structural studies, and many factors can affect the likelihood of obtaining diffraction-quality crystals. The daunting prospect for almost any new membrane target is that every stage, including expression, solubilization, purification and crystallization, will prove challenging. In this light, it is encouraging that remarkably similar techniques have been employed for many successfully crystallized membrane targets, including, for example, the observation that most structures have been solved using just a handful of different detergents throughout the isolation and crystallization steps. It is equally significant that simple two-step purification strategies predominate. Clearly, new methods for the production of membrane proteins will continue to be developed. Nevertheless, for a field that is quite likely to be entering an acute phase of growth, a solid foundation has been established for reasonable 'first-try' strategies.

References

Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. (2002). *Crystal structure of Escherichia coli MscS, a voltage-modulated and mechanosensitive channel. Science*, **298**, 1582–1587.

Berg, B. van den, Clemons, W. M., Collinson, I., Modis, Y., Hartmann, E., Harrison, S. C. & Rapoport, T. A. (2004). *X-ray structure of a protein-conducting channel. Nature (London)*, **427**, 36–44.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *The Protein Data Bank. Acta Cryst.* D**58**, 899–907.

Carpenter, E. P., Beis, K., Cameron, A. D. & Iwata, S. (2008). *Overcoming the challenges of membrane protein crystallography. Curr. Opin. Struct. Biol.* **18**, 581–586.

Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. (2007). *His-tag impact on structure. Acta Cryst.* D**63**, 295–301.

Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. (1998). *Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. Science*, **282**, 2220–2226.

Chen, Y. J., Pornillos, O., Lieu, S., Ma, C., Chen, A. P. & Chang, G. (2007). *X-ray structure of EmrE supports dual topology model. Proc. Natl Acad. Sci. USA*, **104**, 18999–19004.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K. & Stevens, R. C. (2007). *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science*, **318**, 1258–1265.

Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J. R., Booth, V., Mackereth, C. D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K. L., Wu, N., McIntosh, L. P., Gehring, K., Kennedy, M. A., Davidson, A. R., Pai, E. F., Gerstein, M., Edwards, A. M. & Arrowsmith, C. H. (2000). *Structural proteomics of an archaeon. Nat. Struct. Biol.* **7**, 903–909.

Cortes, D. M. & Perozo, E. (1997). *Structural dynamics of the Streptomyces lividans K+ channel (SKC1): oligomeric stoichiometry and stability. Biochemistry*, **36**, 10343–10352.

Day, P. W., Rasmussen, S. G., Parnot, C., Fung, J. J., Masood, A., Kobilka, T. S., Yao, X. J., Choi, H. J., Weis, W. I., Rohrer, D. K. & Kobilka, B. K. (2007). *A monoclonal antibody for G protein-coupled receptor crystallography. Nat. Methods*, **4**, 927–929.

Dobrovetsky, E., Lu, M. L., Andorn-Broza, R., Khutoreskaya, G., Bray, J. E., Savchenko, A., Arrowsmith, C. H., Edwards, A. M. & Koth, C. M. (2005). *High-throughput production of prokaryotic membrane proteins. J. Struct. Funct. Genomics*, **6**, 33–50.

Dobrovetsky, E., Menendez, J., Edwards, A. M. & Koth, C. M. (2007). *A robust purification strategy to accelerate membrane proteomics. Methods*, **41**, 381–387.

Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). *The structure of the* *potassium channel: molecular basis of K+ conduction and selectivity. Science*, **280**, 69–77.

Engel, C. K., Chen, L. & Privé, G. G. (2002). *Stability of the lactose permease in detergent solutions. Biochim. Biophys. Acta*, **1564**, 47–56.

Eshaghi, S., Hedrén, M., Nasser, M. I., Hammarberg, T., Thornell, A. & Nordlund, P. (2005). *An efficient strategy for high-throughput expression screening of recombinant integral membrane proteins. Protein Sci.* **14**, 676–683.

Geertsma, E. R., Groeneveld, M., Slotboom, D. J. & Poolman, B. (2008). *Quality control of overexpressed membrane proteins. Proc. Natl Acad. Sci. USA*, **105**, 5722–5727.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B. M., Bray, J., Gileadi, O., Knapp, S., Oppermann, U., Arrowsmith, C., Hui, R., Ming, J., dhe-Paganon, S., Park, H. W., Savchenko, A., Yee, A., Edwards, A., Vincentelli, R., Cambillau, C., Kim, R., Kim, S. H., Rao, Z., Shi, Y., Terwilliger, T. C., Kim, C. Y., Hung, L. W., Waldo, G. S., Peleg, Y., Albeck, S., Unger, T., Dym, O., Prilusky, J., Sussman, J. L., Stevens, R. C., Lesley, S. A., Wilson, I. A., Joachimiak, A., Collart, F., Dementieva, I., Donnelly, M. I., Eschenfeldt, W. H., Kim, Y., Stols, L., Wu, R., Zhou, M., Burley, S. K., Emtage, J. S., Sauder, J. M., Thompson, D., Bain, K., Luz, J., Gheyi, T., Zhang, F., Atwell, S., Almo, S. C., Bonanno, J. B., Fiser, A., Swaminathan, S., Studier, F. W., Chance, M. R., Sali, A., Acton, T. B., Xiao, R., Zhao, L., Ma, L. C., Hunt, J. F., Tong, L., Cunningham, K., Inouye, M., Anderson, S., Janjua, H., Shastry, R., Ho, C. K., Wang, D., Wang, H., Jiang, M., Montelione, G. T., Stuart, D. I., Owens, R. J., Daenke, S., Schütz, A., Heinemann, U., Yokoyama, S., Büssow, K. & Gunsalus, K. C. (2008). *Protein production and purification. Nat. Methods*, **5**, 135–146.

Grisshammer, R. & Tate, C. G. (1995). *Overexpression of integral membrane proteins for structural studies. Q. Rev. Biophys.* **28**, 315–422.

Guzman, L. M., Belin, D., Carson, M. J. & Beckwith, J. (1995). *Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. J. Bacteriol.* **177**, 4121–4130.

Hanson, M. A., Cherezov, V., Griffith, M. T., Roth, C. B., Jaakola, V. P., Chien, E. Y., Velasquez, J., Kuhn, P. & Stevens, R. C. (2008). *A specific cholesterol binding site is established by the 2.8 Å structure of the human beta2-adrenergic receptor. Structure*, **16**, 897–905.

Helenius, A., McCaslin, D. R., Fries, E. & Tanford, C. (1979). *Properties of detergents. Methods Enzymol.* **56**, 734–749.

Hjelmeland, L. M. (1990). *Removal of detergents from membrane proteins. Methods Enzymol.* **182**, 277–282.

Hunte, C. & Michel, H. (2002). *Crystallisation of membrane proteins mediated by antibody fragments. Curr. Opin. Struct. Biol.* **12**, 503–508.

Ishihara, G., Goto, M., Saeki, M., Ito, K., Hori, T., Kigawa, T., Shirouzu, M. & Yokoyama, S. (2005). *Expression of G protein coupled receptors in a cell-free translational system using detergents and thioredoxin-fusion vectors. Protein Expr. Purif.* **41**, 27–37.

Jaakola, V. P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P. & Stevens, R. C. (2008). *The 2.6 Å crystal structure of a human A2A adenosine receptor bound to an antagonist. Science*, **322**, 1211–1217.

Klammt, C., Schwarz, D., Eifler, N., Engel, A., Piehler, J., Haase, W., Hahn, S., Dötsch, V. & Bernhard, F. (2007). *Cell-free production of G protein-coupled receptors for functional and structural studies. J. Struct. Biol.* **158**, 482–493.

Krautwurst, D., Yau, K. W. & Reed, R. R. (1998). *Identification of ligands for olfactory receptors by functional expression of a receptor library. Cell*, **95**, 917–926.

Kuruma, Y., Nishiyama, K., Shimizu, Y., Müller, M. & Ueda, T. (2005). *Development of a minimal cell-free translation system for the synthesis of presecretory and integral membrane proteins. Biotechnol. Prog.* **21**, 1243–1251.

Lemieux, M. J., Song, J., Kim, M. J., Huang, Y., Villa, A., Auer, M., Li, X. D. & Wang, D. N. (2003). *Three-dimensional crystallization of the Escherichia coli glycerol-3-phosphate transporter: a member of the major facilitator superfamily. Protein Sci.* **12**, 2748–2756.

Lewinson, O., Lee, A. T. & Rees, D. C. (2008). *The funnel approach to the precrystallization production of membrane proteins. J. Mol. Biol.* **377**, 62–73.

Lichty, J. J., Malecki, J. L., Agnew, H. D., Michelson-Horowitz, D. J. & Tan, S. (2005). *Comparison of affinity tags for protein purification. Protein Expr. Purif.* **41**, 98–105.

Long, S. B., Campbell, E. B. & Mackinnon, R. (2005). *Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. Science*, **309**, 897–903.

Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. (2007). *Atomic structure of a voltage-dependent K+ channel in a lipid membrane-like environment. Nature (London)*, **450**, 376–382.

Lunin, V. V., Dobrovetsky, E., Khutoreskaya, G., Zhang, R., Joachimiak, A., Doyle, D. A., Bochkarev, A., Maguire, M. E., Edwards, A. M. & Koth, C. M. (2006). *Crystal structure of the CorA Mg2+ transporter. Nature (London)*, **440**, 833–837.

Magnani, F., Shibata, Y., Serrano-Vega, M. J. & Tate, C. G. (2008). *Co-evolving stability and conformational homogeneity of the human adenosine A2a receptor. Proc. Natl Acad. Sci. USA*, **105**, 10744–10749.

Miroux, B. & Walker, J. E. (1996). *Over-production of proteins in Escherichia coli: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels. J. Mol. Biol.* **260**, 289–298.

Neugebauer, J. M. (1990). *Detergents: an overview. Methods Enzymol.* **182**, 239–253.

Newby, Z. E., O'Connell, J. D. III, Gruswitz, F., Hays, F. A., Harries, W. E., Harwood, I. M., Ho, J. D., Lee, J. K., Savage, D. F., Miercke, L. J. & Stroud, R. M. (2009). *A general protocol for the crystallization of membrane proteins for X-ray structural investigation. Nat. Protoc.* **4**, 619–637.

Newstead, S., Ferrandon, S. & Iwata, S. (2008). *Rationalizing alpha-helical membrane protein crystallization. Protein Sci.* **17**, 466–472.

Payandeh, J. & Pai, E. F. (2006). *A structural basis for Mg2+ homeostasis and the CorA translocation cycle. EMBO J.* **25**, 3762–3773.

Prive, G. G. (2007). *Detergents for the stabilization and crystallization of membrane proteins. Methods*, **41**, 388–397.

Raman, P., Cherezov, V. & Caffrey, M. (2006). *The Membrane Protein Data Bank. Cell. Mol. Life Sci.* **63**, 36–51.

Rasmussen, S. G., Choi, H. J., Rosenbaum, D. M., Kobilka, T. S., Thian, F. S., Edwards, P. C., Burghammer, M., Ratnala, V. R., Sanishvili, R., Fischetti, R. F., Schertler, G. F., Weis, W. I. & Kobilka, B. K. (2007). *Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. Nature (London)*, **450**, 383–387.

Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Yao, X. J., Weis, W. I., Stevens, R. C. & Kobilka, B. K. (2007). *GPCR engineering yields high-resolution structural insights into beta2-adrenergic receptor function. Science*, **318**, 1266–1273.

Sakai, H. & Tsukihara, T. (1998). *Structures of membrane proteins determined at atomic resolution. J. Biochem.* **124**, 1051–1059.

Savage, D. F., Anderson, C. L., Robles-Colmenares, Y., Newby, Z. E. & Stroud, R. M. (2007). *Cell-free complements in vivo expression of the E. coli membrane proteome. Protein Sci.* **16**, 966–976.

Schwarz, D., Dötsch, V. & Bernhard, F. (2008). *Production of membrane proteins using cell-free expression systems. Proteomics*, **8**, 3933–3946.

Schwarz, D., Klammt, C., Koglin, A., Lohr, F., Schneider, B., Dotsch, V. & Bernhard, F. (2007). *Preparative scale cell-free expression systems: new tools for the large scale preparation of integral membrane proteins for functional and structural studies. Methods*, **41**, 355–369.

Serrano-Vega, M. J., Magnani, F., Shibata, Y. & Tate, C. G. (2008). *Conformational thermostabilization of the beta1-adrenergic receptor in a detergent-resistant form. Proc. Natl Acad. Sci. USA*, **105**, 877–882.

Studier, F. W. & Moffatt, B. A. (1986). *Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J. Mol. Biol.* **189**, 113–130.

Sukharev, S. I., Schroeder, M. J. & McCaslin, D. R. (1999). *Stoichiometry of the large conductance bacterial mechanosensitive channel of E. coli. A biochemical study. J. Membr. Biol.* **171**, 183–193.

Surade, S., Klein, M., Stolt-Bergner, P. C., Muenke, C., Roy, A. & Michel, H. (2006). *Comparative analysis and 'expression space' coverage of the production of prokaryotic membrane proteins for structural genomics. Protein Sci.* **15**, 2178–2189.

Wagner, S., Klepsch, M. M., Schlegel, S., Appel, A., Draheim, R., Tarry, M., Högbom, M., van Wijk, K. J., Slotboom, D. J., Persson, J. O. & de Gier, J. W. (2008). *Tuning Escherichia coli for membrane protein overexpression. Proc. Natl Acad. Sci. USA*, **105**, 14371–14376.

Wang, D. N., Safferling, M., Lemieux, M. J., Griffith, H., Chen, Y. & Li, X. D. (2003). *Practical aspects of overexpressing bacterial secondary membrane transporters for structural studies. Biochim. Biophys. Acta*, **1610**, 23–36.

Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G., Tate, C. G. & Schertler,

G. F. (2008). *Structure of a beta1-adrenergic G-protein-coupled receptor. Nature (London)*, **454**, 486–491.

Warne, T., Serrano-Vega, M. J., Tate, C. G. & Schertler, G. F. (2009). *Development and crystallization of a minimal thermostabilised G protein-coupled receptor. Protein Expr. Purif.* **65**, 204–213.

Wiener, M. C. (2004). *A pedestrian guide to membrane protein crystallization. Methods*, **34**, 364–372.

Willis, M. S. & Koth, C. M. (2008). *Structural proteomics of membrane proteins: a survey of published techniques and design of a rational high-throughput strategy. Methods Mol. Biol.* **426**, 277–295.

International Tables for Crystallography (2012). Vol. F, Chapter 4.1, pp. 99–121.

PART 4. CRYSTALLIZATION

Chapter 4.1. General methods

C. Sauter, B. Lorber, A. McPherson and R. Giegé

4.1.1. Introduction

4.1.1.1. Prologue

Macromolecular crystals are much softer than salt crystals. They contain ∼50% of solvent on average, with values ranging from as little as 25 up to 90%. A direct consequence is that, although morphologically indistinguishable, crystals of macromolecules differ in many respects from crystals of low-molecular-mass compounds. While the latter exhibit firm lattice forces, are highly ordered, generally physically hard and brittle, easy to manipulate, can usually be exposed to air, have strong optical properties and diffract X-rays intensely, crystals of macromolecules are, by comparison, smaller in size, they crush easily, disintegrate if allowed to dehydrate, exhibit weak optical properties and diffract X-rays poorly. They are temperature sensitive and undergo extensive damage after prolonged exposure to radiation.

Proteins or nucleic acids build up a crystalline scaffold, which may be imagined as an ordered gel with extensive interstitial spaces through which small molecules can diffuse freely. In proportion to molecular mass, large macromolecules establish far fewer packing interactions than do small molecules inside crystalline lattices. Since these contacts are responsible for the integrity of the crystals, this largely explains the differences in properties between the two types of crystals. Thus, liquid channels and solvent cavities are directly responsible for the generally poor diffraction properties of macromolecular crystals. Owing to the large spaces between adjacent molecules and the related weak lattice forces, every molecule in the crystal may not occupy exactly equivalent orientations and positions. Furthermore, because of their structural complexity and their conformational dynamics, macromolecules in a given crystal form may exhibit slight variations in their folding patterns or dispositions of side chains.

However, high solvent content is not as negative as it might appear at first glance. It allows maintenance of the macromolecular structures virtually unchanged from those in bulk solvent. As a consequence, ligand binding, enzymatic and spectroscopic characteristics, and other biochemical features are essentially the same as for the native molecule in solution. In addition, the dimensions of solvent channels are such that conventional chemical compounds, such as ions and heavy atoms, substrates or other ligands, may be freely diffused into and out of the crystals. Thus, many crystalline enzymes, though immobilized, are completely accessible for experimentation through alteration of the surrounding mother liquor.

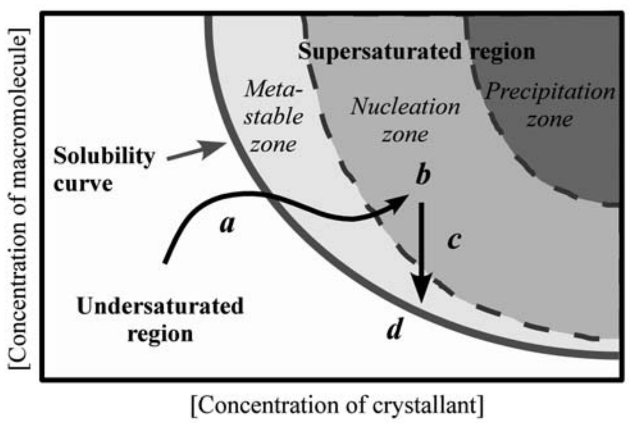
The intrinsic instability of most macromolecules requires that conditions suitable for crystal growth are those that do not perturb their molecular properties. This explains why crystals must be grown from solutions compatible with the target macromolecules, *i.e.* within a narrow range of pH, temperature or ionic strength. Finally, because hydration is essential for the maintenance of structure, crystals of macromolecules must be always bathed in the mother liquor, even during data collection (except in the practice of cryocrystallography).

4.1.1.2. Crystallization principles

The crystallization of biological macromolecules has often been considered unpredictable, although it obeys the same principles as that of small molecules (Giegé *et al.*, 1995; McPherson *et al.*, 1995; Rosenberger, 1996; Chernov, 2003; McPherson & Giegé, 2007). It is, similarly, a multiparametric process. The difference compared with conventional crystal growth arises from the biochemical and biophysical properties of proteins or nucleic acids, and crystallization methods must take into account these features. The methods described below apply for most proteins, large RNAs, multimacromolecular complexes and viruses (for small oligonucleotides or peptides, crystallization by dialysis is not appropriate). For hydrophobic membrane proteins, special techniques are required (see Chapter 4.2).

Crystallization proceeds from macromolecules in solution that ‘aggregate’ upon entering a supersaturated state and eventually undergo a phase transition. This leads to nuclei formation and ultimately to crystals that grow by different mechanisms. Supersaturation is the driving force of crystallization and is defined as the ratio $[C]/[S]$, where $[C]$ and $[S]$ are the initial concentration of the macromolecule and its final concentration at saturation, *i.e.* its solubility. Nucleation is homogeneous when nuclei form in the bulk of the solution, but heterogeneous when they preferentially form on walls of crystallization vessels, on solid particles (dust, aggregates, seeds), or on the surface of existing crystals. Unlike most conventional crystals, protein crystals are, in general, not initiated from seeds, but are nucleated *ab initio* at high levels of supersaturation that can reach 200 to 1000% (in what follows and for simplicity, the generic name ‘protein’ is used for macromolecule). It is this high degree of supersaturation that, in large part, distinguishes protein crystal formation from that of conventional crystals. That is, once a stable nucleus has formed, it subsequently grows under very unfavourable conditions of excessive supersaturation. Distant from the metastable zone, where ordered growth could occur, crystals rapidly accumulate nutrient molecules, as well as impurities. They also concomitantly accumulate statistical disorder and a high frequency of defects that exceeds those observed for most conventional crystals.

The different stages of crystallization (*i.e.* pre-nucleation, nucleation, growth, cessation of growth) can be visualized in a phase diagram (Fig. 4.1.1.1). In short, phase diagrams are divided into undersaturated regions (where proteins are soluble) and supersaturated regions (where protein crystals nucleate and grow) delimited by the solubility curve. The supersaturated region is thermodynamically out of equilibrium and can be divided into three kinetically dependent domains: a precipitation domain (at extreme supersaturation) where macromolecules rapidly separate from solution in a solid state either amorphous or microcrystalline, a domain (at lower supersaturation) where



Biochemical parameters
Purity of the macromolecule
Ligands, inhibitors, effectors
Aggregation state of the macromolecule
Post-translational modifications
Source of macromolecule
Proteolysis or hydrolysis
Chemical modifications
Genetic variants
Inherent symmetry of the macromolecule
Stability of the macromolecule
Isoelectric point
History of the sample

Chemical parameters
pH
Crystallant type
Crystallant concentration
Ionic strength
Specific ion effects
Supersaturation
Reductive or oxidative environment
Concentration of the macromolecule
Metal ions
Crosslinkers or polyions
Detergents, surfactants or amphophiles
Non-macromolecular impurities

Physical parameters
Temperature variation
Contact surfaces
Methodology or approach to equilibrium
Gravity
Pressure
Time
Vibrations, sound or mechanical perturbations
Electrostatic or magnetic fields
Dielectric properties of the medium
Viscosity of the medium
Rate of equilibration
Homogeneous or heterogeneous nucleants

Figure 4.1.1.1

Crystallization is a multiparametric process under the control of a great variety of biochemical, chemical and physical parameters. The crystal grower can play with them to drive a macromolecule from the soluble towards the crystalline state. In the top left, a typical two-dimensional phase diagram illustrates how this can be achieved. (*a*) By modifying two parameters (concentrations of macromolecule and crystallant), the system will move from the undersaturated region where the macromolecule is soluble into the supersaturated region beyond the solubility curve where it will try to escape from the solution. The prenucleation trajectory essentially depends on the crystallization method (see Fig. 4.1.3.1). (*b*) As soon as they enter the supersaturation region, macromolecules will tend to aggregate. However, the system needs to cross an energy barrier to produce stable molecular assemblies, the nuclei, and this only happens in the so-called nucleation zone. (*c*) When stable nuclei are formed, they will capture more macromolecular entities from the mother liquor and produce three-dimensional crystals. (*d*) Crystals will grow until the system comes back to the solubility curve, crossing the metastable zone. Growth stops and crystals are in dynamic equilibrium with the mother liquor. When the system is driven to high supersaturation, macromolecules may rather produce amorphous or microcrystalline precipitates than useful monocrystals. Note that the limits between zones of the supersaturated zone (dashed curve) move with time.

nucleation occurs spontaneously and a metastable domain (at low supersaturation) where nucleation does not occur spontaneously but where crystals grow. This domain is favourable for seeding. The wisdom of the crystal grower will be to take advantage of an overall understanding of phase diagrams for designing crystallization strategies and selecting favourable solvent conditions (Sauter, Lorber *et al.*, 1999; Asherie, 2004).

4.1.2. Main parameters that affect crystallization of macromolecules

4.1.2.1. Crystallizing agents

Crystallizing agents (also referred to as crystallants or precipitants) are defined as those chemical species that are able to precipitate a soluble macromolecule and hence to bring it into a supersaturated state needed for its crystallization. They fall into four categories: (i) salts, (ii) organic solvents, (iii) long-chain polymers, and (iv) low-molecular-mass polymers and non-volatile organic compounds (McPherson, 1990). The first two classes are typified by ammonium sulfate and ethanol; higher polymers such as PEGs (polyethylene glycols) of molecular weight (M_r) > 1000, are characteristic of the third. In the fourth are placed compounds such as MPD (2-methyl-2,4-pentanediol) and low-molecular-mass PEGs.

Salts. Salts exert their effects by dehydrating proteins through competition for water molecules (Green & Hughes, 1955). Their ability to do this is roughly proportional to the square of the valences of the ionic species composing the salt. Thus, multivalent ions, particularly anions, are the most efficient. One might think there would be little variation between different salts, so long as their ionic valences were the same, or between two different sulfates, such as lithium and ammonium sulfates. This, however, is often not the case. In addition to salting out, which is a dehydration effect, or lowering of the chemical activity of water, there are specific protein–ion interactions that have other consequences (Riès-Kautt & Ducruix, 1991; Ducruix & Giegé, 1999). This is true because of the polyvalent character of individual proteins, their structural complexity, and the dependence of their physical properties on environmental conditions and interacting molecules. Therefore, restricting the search to one or two salts is never sufficient when attempting to crystallize a protein. Changing the salt may produce crystals of varied quality, morphology and diffraction properties.

It is usually not possible to predict the molarity of a salt required to crystallize a particular protein without some prior knowledge of its behaviour. In general, the concentration is just a small percentage less than that yielding an amorphous precipitate. To determine the precipitation point with a given crystallant, a 10 µl droplet of a 5–15 mg ml^{-1} protein solution is placed in the well of a depression slide and observed under a microscope as increasing amounts of salt solution or organic solvent (in 1–2 µl increments) are added. If the well is sealed between additions with a coverslip, the increases can be made over a period of many hours. Indeed, the droplet should equilibrate 10–30 min after each addition, and longer in the neighbourhood of the precipitation point.

Widely used and the most successful salt to grow protein crystals is ammonium sulfate. A variety of other salts included in crystallization screens yield success as well. Sodium malonate, not systematically present in screens, deserves particular attention, since it gave the highest success rate (better than ammonium sulfate) in a comparative study designed to identify the most effective salts for protein crystallization (McPherson, 2001). Following this example, other organic salts have been added to the repertoire (McPherson & Cudney, 2006).

Organic solvents. The most common organic solvents (volatile compounds) utilized are ethanol, methanol, isopropanol, acetone and dioxane. They have been frequently used for crystallizing nucleic acids, particularly tRNAs and duplex oligonucleotides

(Dock *et al.*, 1984). This, in part, stems from the greater tolerance of polynucleotides to organic solvents and their polyanionic character, which appears more sensitive to dielectric effects than proteins. Volatile organic solvents should be used at low temperature and should be added slowly and with good mixing.

Long-chain polymers. PEGs are polymers of various lengths that are useful in crystallogenesis (McPherson, 1976). The larger species with $M_r > 1000$ exist at room temperature as either waxy solids or powders. The M_r's specified by manufacturers are mean values, and the distribution around these means varies. In addition to volume exclusion properties, PEGs share characteristics with salts that compete for water and produce dehydration, and with organic solvents that reduce the dielectric properties of the medium. PEGs also have the advantage of being effective at minimal ionic strength and providing low-electron-density media. The first feature is important because it leads to stronger affinities for ligands than high-ionic-strength media. Thus, there is greater ease in obtaining heavy-atom derivatives and in forming protein–ligand complexes. The second characteristic, their low electron density, implies a lower noise level for structures derived by X-ray diffraction.

The most useful PEGs in crystallogenesis are those in the range 2000–8000. Sizes are generally not completely interchangeable for a given protein, and this parameter has to be optimized by empirical means. An advantage of PEG over other crystallants is that most proteins crystallize within a rather narrow range of PEG concentrations (~4–18%). In addition, the PEG concentration at which crystals form is rather insensitive, and if one is within 2–3% of the optimal value, some success will be achieved. The advantage is that, when conducting initial trials, one can use a fairly coarse selection of concentrations. This means fewer trials with a corresponding reduction in the amount of material expended. Note that PEGs are often used in conjunction with salts.

Since PEG is not volatile, this agent must be used like salt and equilibrated with the protein by dialysis, slow mixing or vapour diffusion. This latter approach is the most popular. When the reservoir concentration is in the range 5–12%, the protein solution to be equilibrated should initially be at about half that concentration. When the target PEG concentration is higher than 12%, it is advisable to initiate the equilibration at no more than 4–5% below the final value. This reduces lags in time during which the protein might denature. Crystallization of proteins with PEG has proved most successful when ionic strength is low, and more difficult when high. If crystallization proceeds too rapidly, addition of some neutral salt may be used to slow growth. PEG can be used over the entire pH range and a broad temperature range. It should be noted that solutions with PEG may serve as media for microbes, particularly moulds, and if crystallization is attempted at room temperature or over extended periods of time, then retardants, such as azide (~0.1%), must be included in the protein solutions.

Low-molecular-mass polymers and non-volatile organic compounds. Low-molecular-mass PEG species with $M_r < 1000$, MPD and glycerol are widely used. These are oily liquids that increase the viscosity of the crystallization samples. Glycerol and other polyols have structure-ordering potential and might be useful for crystallizing flexible proteins (Sousa, 1995). MPD has properties midway between PEG and organic solvents, which makes it one of the most successful crystallants for promoting crystallization of biological macromolecules (Anand *et al.*, 2002). Besides efficacy with proteins, ~50% of oligonucleotides were crystallized with MPD (Li *et al.*, 2007). These low-molecular-mass compounds are often used in combination with salts or long-chain PEGs.

Recent data show that room-temperature ionic liquids have the potential for macromolecular crystallogenesis (Pusey *et al.*, 2007) and can be added to the list of crystallants. They are organic salts that can solubilize proteins while preserving structural stability and enzymatic activity. With regard to crystallization applications, it is worth noticing that imidazolium-based ionic liquids enhance protein folding and suppress aggregation (Lange *et al.*, 2005). As a promising result, several ionic liquids used in the range 3–30%(v/v) yielded crystals of trypsin and lysozyme (Judge *et al.*, 2009).

4.1.2.2. Physical, physical–chemical and biochemical variables

Many physical, chemical and biological variables influence, to a greater or lesser extent, the crystallization of macromolecules (Fig. 4.1.1.1). The difficulty in assigning the relative importance for each factor is substantial for several reasons. First, each macromolecule has different surface properties and this even applies to molecules that differ by no more than one or a few residues. Therefore one can hardly predict in advance the specific values of a variable or sets of conditions that might be most profitably explored. Furthermore, crystallization parameters are not independent and their interrelations may be difficult to discern. Thus, it is not easy to elaborate firm guidelines related to physical or chemical factors that can increase the success in crystallizing a particular macromolecule. Among them, only temperature and pH have been studied carefully. For pressure or magnetic and electric fields, few investigations have been carried out and virtually nothing is known on the effects of sound, vibrations or viscosity on the growth or final quality of protein crystals (see Section 4.1.4.2).

Current physical–chemical variables. Temperature may be of great importance (Judge *et al.*, 1999; Astier & Veesler, 2008; Lin *et al.*, 2008). In general, it is wise to conduct parallel investigations at 4 and 20 °C. Even if no crystals are observed at either temperature, differences in the solubility behaviour of a protein with various crystallants and effector molecules may give some indication as to whether temperature is likely to play an important role (Christopher *et al.*, 1998). Generally, the solubility of a protein is more sensitive to temperature at low ionic strength than high. One must remember, however, that diffusion rates are less, and equilibration occurs more slowly at colder than higher temperature, so that the time required for crystal formation may be longer at colder temperatures. Although most crystallization trials are done at low (~4° C) or medium (~20 °C) temperatures, warmer temperatures in the range 35–40 °C should not be ignored, particularly with molecules that tend to aggregate and with nucleic acids (Dock-Bregeon *et al.*, 1988).

Another important variable is pH (Judge *et al.*, 2009). This follows since the charge character of a macromolecule and all of its consequences are intimately dependent on the ionization state of its components. Not only does its net charge change with pH (and the charge distribution), but also its dipole moment, conformation and often its aggregation state. Thus, an investigation of the behaviour of a specific macromolecule as a function of pH is an essential analysis that should be carried out in performing crystallization assays. Various buffer systems that are compatible with macromolecules are available for that purpose (Newman, 2004). As with temperature, the procedure is, first, to conduct trials at coarse intervals over a broad pH range and then to refine trials in the neighbourhood of those that showed promise. In refining the pH for optimal growth, it should be recalled that the difference between amorphous precipitate, microcrystals and large single crystals might be due to a ΔpH of less than half a unit.

Other physical variables. Convection and diffusion are two correlated parameters that govern mass transport during the crystallization process, and studies on the physics of macromolecular crystal growth, in particular in relation to microgravity research (Giegé *et al.*, 1995; McPherson, 1996; Kundrot *et al.*, 2001; Helliwell & Chayen, 2007), have evaluated their importance and led to advanced methods of crystallization (see Section 4.1.4). On the other hand, mechanical or other types of physical perturbation of the crystallization media, *e.g.* by external electric or ultrasonic fields (Nanev & Penkova, 2001; Kakinouchi *et al.*, 2006), by continuous light irradiation (Veesler *et al.*, 2006) and by laser pulses (Lee *et al.*, 2008), might be used to induce nucleation. Local increase of protein concentration due to laser-induced cavitation could be a trigger of nucleation (Yoshikawa *et al.*, 2009).

Biochemical variables. It is now generally appreciated that, indeed, the protein is itself the most important and influential parameter in the crystallization endeavour (Dale *et al.*, 2003). When conventional approaches to crystallization have been exhausted there are two remaining options: further purifying the protein, or modifying the protein. The latter may be accomplished by genetic means using recombinant DNA techniques; that is, single or multiple point mutations may be introduced, or truncated forms of the polypeptide generated (see Section 4.1.2.4). Alternatively, traditional chemical reactions may be used to modify existing amino acids, or exposure to modifying enzymes may be employed, for example, the production of truncations by limited proteolysis. There are many examples of truncated proteins being successfully crystallized when the full-length polypeptide could not, *e.g.*, with many aminoacyl-tRNA synthetases (Giegé *et al.*, 2008). With the predictive capability of modern amino-acid sequence analysis and mass spectrometry to identify domains within proteins, designed constructs of predetermined lengths are becoming increasingly used. It has also been suggested that some surface amino acids, such as lysine and glutamic acid, inflict entropic costs when a protein crystallizes (Derewenda, 2004). Modification of such residues by reductive methylation appears to be a useful approach (Kim *et al.*, 2008) that in some cases enhances crystallizability. For instance, the addition of two methyl groups on the side chain of nine lysine residues of a nuclease introduced 44 cohesive CH intramolecular contacts and resulted in crystals diffracting at 1.2 Å resolution (Shaw *et al.*, 2007).

Histidine tags and proteins conjugated with a second protein, such as the maltose binding protein, are frequently produced as a basis for purification procedures for recombinant proteins. There is no reason why these ‘tagged’ or conjugated proteins should not crystallize and, indeed, many of them do. On the other hand, removal of the tag or conjugate and subsequent crystallization trials provide a further opportunity to obtain crystals of a protein, and should certainly be tried (Carson *et al.*, 2007).

Another possibility to gain access to the structure of proteins recalcitrant to crystallize is to try crystallization of their orthologues. Indeed, proteins are adapted to their cellular environments and orthologues of the same physiological function can exhibit structural and physical–chemical idiosyncrasies leading to modified crystallizability. Thus proteins and macromolecular assemblages from extremophile organisms often crystallize more readily than their cousins from mesophiles or from higher eukaryotes, as shown for instance with aminoacyl-tRNA synthetases (Giegé *et al.*, 2008).

In nucleic acid crystallization, the large size of natural DNA and of many RNA molecules necessitates the design and preparation of adequate structural modules or domains amenable to crystal production. Tailoring an adequate module is often the prerequisite for its crystallization (Ducruix & Giegé, 1999; Holbrook *et al.*, 2001; Golden & Kundrot, 2003). Since RNA architecture is very modular, specific protein-binding domains can be introduced (typically small RNA hairpins) in the target RNA sequence to promote the association with a protein partner (*e.g.* the spliceosomal protein U1A) and hence to increase the possibilities of making packing interactions (Ferré-D’Amaré, 2010). Similarly, this engineering approach can be applied to proteins. A typical example is the production of crystals diffracting to high resolution of a G-protein-coupled receptor. In that case one flexible loop of this membrane protein was replaced by the sequence of T4 lysozyme and this globular exogenic domain created new contacts leading to three-dimensional packing (Cherezov *et al.*, 2007). Another example of engineering is the introduction of a leucine half-zipper into an aminoacyl-tRNA synthetase that has engendered multiple high-quality crystals (Guo *et al.*, 2010).

4.1.2.3. *Additives*

Certain chemical compounds or small molecules may have dramatic effects on the success of protein crystallization. Additives, as they are often called (McPherson, 1982, 1999; Sauter, Ng *et al.*, 1999), can be decisive in macromolecular crystallization. The most commonly used type of additives, and the only class for which we have any rational basis, are those which may, for physiological reasons, be bound by the protein with consequent favourable changes in its physical–chemical properties or conformation. These include coenzymes and prosthetic groups, inhibitors, enzymatic products, ions and other effector molecules. Often the liganded form is structurally defined and stable, while the unliganded form is not, which will lead to the crystallization of the former and not of the latter. Polyamines and metal ions are useful for nucleic acids (Dock *et al.*, 1984; Ducruix & Giegé, 1999). In terms of crystallization, complexes have to be treated as almost entirely separate problems. This may permit a new opportunity for growing crystals if the native molecule is obstinate. Just as natural substrates or inhibitors are often useful, they can also have the opposite effect of obstructing crystal formation. In such cases, care must be taken to eliminate them from the mother liquor and from the purified protein before crystallization is attempted.

Numerous cases have, however, been reported where molecules were observed to make crucial interactions between macromolecules in the crystal that either helped guide or secure formation of the lattice. Such molecules sometimes had a physiological basis for their unexpected presence, but frequently not. They simply provided essential or at least helpful cross-links within the crystal. Additives that are used in protein crystallization or that might be appropriate for use in crystallization may be classified into nine categories.

Small physiological ligands. Physiologically or biochemically relevant small molecules such as coenzymes, substrate analogues, inhibitors, metal ions, prosthetic groups *etc.* are the first. These bind at the active sites of enzymes, or at specific sites elsewhere on protein molecules, and may promote more stable, homogeneous conformations, or they may induce conformational changes into alternate states. In any case, the ultimate protein–ligand complex may exhibit a more monodisperse, less dynamic character. The pertinent molecules are, here, specific to the individual protein under study, and their selection for inclusion in mother liquors is amenable to rational analysis informed by the enzymology and biochemistry of the protein under study. That is, one considers all of the possible ligands of the protein and includes them in the screen of potential crystallization conditions.

Chemical protectants. These include reductants, such as 2-mercaptoethanol, gluthathione or dithiothreitol, which prevent oxidation, heavy metal ions, scavengers such as EDTA and EGTA (ethylenediaminetetraacetic acid and ethylene glycol tetraacetic acid), and compounds intended to prevent microbial infection such as sodium azide, pheno- or chlorobutanol. These compounds are generally included for well understood reasons, their effects are predictable and their impact on the crystallization process usually, but not always, of marginal significance. They prevent protein 'ageing', as do protease or nuclease inhibitors (Ducruix & Giegé, 1999), and are desirable when crystallization requires a long period of time to reach completion or when carried out at room temperature in PEG, or in low-ionic-strength solutions (*i.e.* conditions favourable for the growth of contaminating microorganisms).

Solubilizing agents and detergents. These include quaternary ammonium salts (Mirzabekov *et al.*, 1972), sulfobetains (Goldberg *et al.*, 1996), chaotropes like urea (Bolen, 2004), and a range of surfactant and detergent molecules (Neugebauer, 1990; Zulauf, 1990; Wiener, 2004). Because of the interest in membrane proteins, this class of additives was extensively studied and has been broadly applied to many proteins. Remarkably, there is still no consensus on which are most useful, which should be included in screening conditions, or even how they function in the solubilization of macromolecules. Note that ionic liquids at low concentration can enhance protein solubility and were shown to improve crystallization (Chen *et al.*, 2010).

Poisons. Poisons, as they have traditionally been called (McPherson, 1982, 1999), were originally employed to reduce twinning. These are generally low concentrations, 1–5%(*w*/*v*), of common organic solvents. They include compounds such as ethanol, dimethyl sulfoxide, acetone, dioxane, butanol or MPD. Their role in the crystallization process, even after 50 years of use, remains obscure. They probably enhance the solubility of the proteins and slightly reduce the degree of supersaturation in the mother liquor, as well as lower the dielectric constant of the medium, but they may have other effects as well.

Osmolytes, co-solvents and cosmotropes. These are compounds that exert their effects at relatively high concentrations, 1 *M* or more, and include a wide range of molecules such as sucrose, trehalose and other sugars, proline, trimethyl amine *N*-oxide, glycine, betaine, taurine, sarcosine, and a host of others (Washabaugh & Collins, 1986; Jeruzalmi & Steitz, 1997; Bolen, 2004; Collins, 2004). The effect of their inclusion in the mother liquor is usually to stabilize the native conformation of the protein by altering the interaction of the protein's surface with water, or by altering the hydration layer and possibly the structured waters.

Reversible cross-linkers. It has been proposed that the conformations of proteins might be stabilized, and their dynamic character reduced, by providing the proteins with small molecules that could reversibly cross-link charged groups (carboxyl and amino groups) on the protein's surface, or form intramolecular hydrogen-bonding networks using surface polar groups (Maclean *et al.*, 2002). The molecules that have been explored are usually multivalent molecules such as diamino- or dicarboxylic acid-containing molecules, or aliphatic moieties of various lengths carrying some combination of charged groups. It is not known whether the stabilization of proteins by this means is significant enough to affect their crystallization or not. This potential mechanism of altering crystallization behaviour, however, may indeed be pertinent.

Compounds favouring lattice interactions. Classes of compounds useful for stabilizing proteins through non-covalent intramolecular bonds, as described above, may also help create and stabilize protein crystals by interposing themselves between protein molecules and forming intermolecular cross-links (McPherson, 1999; McPherson & Cudney, 2006; Larson *et al.*, 2007, 2008). These cross-bridges may involve purely electrostatic interactions, or they may rely on hydrogen-bonding arrangements as well. The compounds most favourable for forming such 'lattice interactions' are, again, likely to be multivalent charged compounds, but one might expect that their length, or 'reach', would need to be greater, since they would have to extend from one protein molecule to another.

Nucleation and solubility enhancers. A special class of additives would be those materials or compounds that somehow serve to enhance nucleation, including unique surfaces. These may include low concentrations of PEG (Ray & Bracker, 1986), or other polymeric substances such as jeffamine emulsified in solutions of high salt concentration (Kuznetsov *et al.*, 2000, 2001). The microdroplets of the polymeric phase serve to concentrate the protein locally and provide an interface for nucleation to occur. This category should probably also include things like the gel used in cubic lipidic phase crystallization (Nollert, 2004; Caffrey & Cherezov, 2009) and surfaces that promote epitaxy and heterogeneous nucleation (McPherson & Shlichta, 1988*b*; Chayen *et al.*, 2006; Thakur *et al.*, 2008; Saridakis & Chayen, 2009).

Nucleation enhancers when mixed with current crystallants can also enhance protein solubility (at solute/crystal equilibrium). This is the case, *e.g.*, for acetonitrile, dimethyl sulfoxide, glycerol, MPD and PEGs at relatively high concentrations. As shown with lysozyme and xylose isomerase crystallized, respectively, in the presence of glycerol and acetonitrile, the beneficial effects are nucleation at lower supersaturation and crystals of lower mosaicity (Gosavi *et al.*, 2009).

Ionic liquids. These crystallants (see Section 4.1.2.1) can also be considered as additives. Used in the concentration range 0.05–0.15 *M*, several of them (*e.g.* 1-ethyl-3-methylimidazolium tetrafluoroborate) improved the crystallizability of lysozyme, trypsin and an Fab complex, probably by subtle changes in solution conditions since no ionic liquid ions were observed in the crystal structures (Judge *et al.*, 2009; Chen *et al.*, 2010).

4.1.2.4. *Purity and homogeneity*

The concept of purity is of utmost importance in macromolecular crystallogenesis (Giegé *et al.*, 1986; Rosenberger *et al.*, 1996), even though some macromolecules may crystallize readily from impure solutions (Judge *et al.*, 1998). In practice, macromolecular samples should be cleared of undesired macromolecules and small molecules and, in addition, should be pure in terms of sequence integrity and conformation.

Contaminants can have deleterious effects when they interact with the macromolecule or compete for sites on growing crystals and consequently generate growth disorders (Vekilov & Rosenberger, 1996). Thus, part-per-million amounts of foreign molecules can induce the formation of non-specific aggregates, alter macromolecular solubility, or interfere with nucleation and crystal growth (McPherson *et al.*, 1996, 2004; Moreno *et al.*, 2005; Thomas *et al.*, 1998). These effects are reduced in gel media (Hirschler *et al.*, 1995; Provost & Robert, 1995).

On the other hand, macromolecules that are apparently pure may be microheterogeneous in sequence and/or conformation. Their causes are multiple and can be revealed by analytical tools, such as polyacrylamide gel electrophoresis, isoelectric focusing, nuclear magnetic resonance and mass spectroscopy, but are often overlooked. The most common causes are partial hydrolysis and post-synthetic modifications. Proteolysis represents a major difficulty that must be overcome during protein isolation. In RNAs, hydrolytic cleavages induced by nucleases, metal ions or alkaline pH are common causes of microheterogeneity. These processes can be inhibited by addition of protease or nuclease inhibitors and metal chelators during purification (Ducruix & Giegé, 1999). Heterogeneity in post-synthetic modification patterns in proteins or RNAs can be the result of functional necessity but can also occur when cloned macromolecules are overproduced. Conformational heterogeneity may also originate from ligand binding, intrinsic flexibility of the macromolecule backbones, oxidation of cysteine residues or partial denaturation. Structural homogeneity may be improved by truncation of the flexible parts of the macromolecule under study (see Section 4.1.3). This can also be done *in situ* by addition of trace amounts of protease to crystallization assays (Wernimont & Edwards, 2009). The many reasons that can account for such subtle degradations or modifications explain why altered versions of a macromolecule can be the worst contaminants for its crystallization. Accordingly, the macromolecule itself must be considered as an essential parameter in crystallization (Dale *et al.*, 2003). Control of these phenomena is of crucial importance for the crystallization of macromolecular assemblages, such as the ribosome (Auerbach-Nevo *et al.*, 2005).

Many crystal growers have found a correlation between the outcome of crystallization assays (*i.e.* number of crystals, crystal habit, volume and best diffraction properties) and the quality of macromolecular samples. For this reason, one should never spoil a 'pure' batch by mixing it with another 'pure' one, that may differ as far as microheterogeneities or minute contaminants are concerned. Altogether, purity, good solubility, structural homogeneity and absence of aggregates are good criteria for protein crystallizability (D'Arcy, 1994; Ferré-D'Amaré & Burley, 1997). Dynamic light scattering (DLS) is the appropriate analytical method to verify sample homogeneity, detect aggregates and find solvent conditions that prevent aggregation (Mikol, Hirsch & Giegé, 1990; Borgstahl, 2007; Niesen *et al.*, 2008).

4.1.3. Crystallization arrangements and classical methodologies

Many methods can be used to crystallize macromolecules (Ducruix & Giegé, 1999; McPherson, 1982, 1999). They all aim to bring the macromolecules to an appropriate state of supersaturation (Fig. 4.1.3.1). Although vapour-phase equilibrium and dialysis techniques are favoured, batch and free interface diffusion methods are often used. Besides the physical and chemical variables that affect crystallization (Fig. 4.1.1.1), macromolecular crystal growth is influenced by the crystallization method itself and the geometry of the setup. Generally, in current methods, growth is promoted by the non-equilibrium nature of the crystallization process, which seldom occurs at constant protein concentration. This introduces changes in supersaturation and, hence, may lead to changes in growth mechanism. Crystallization at constant protein concentration, however, can be achieved in special arrangements based on liquid circulation cells (Vekilov & Rosenberger, 1998).

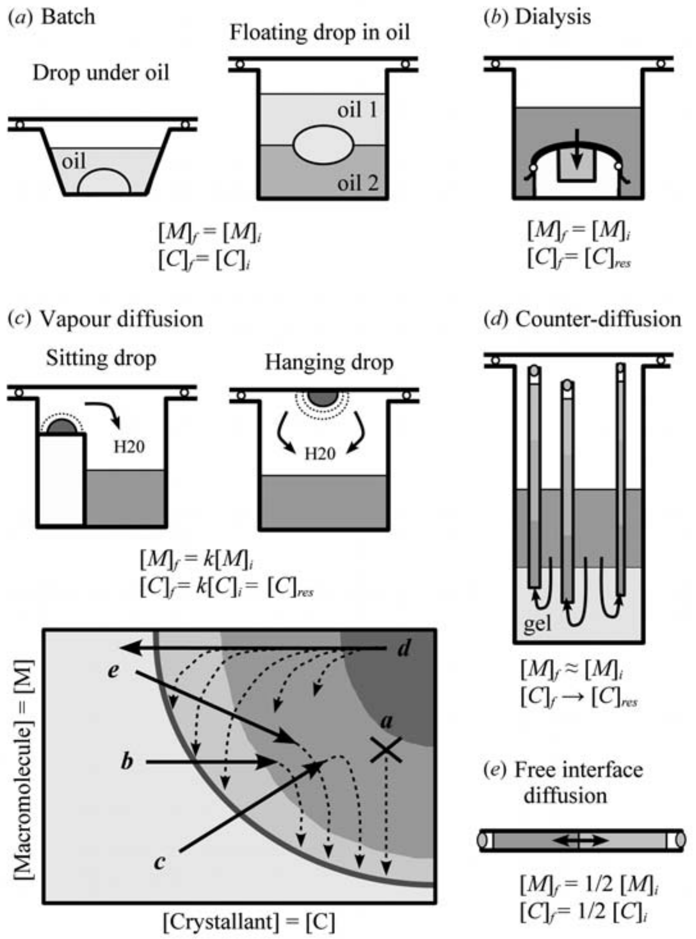


Figure 4.1.3.1

Principles of major methods used to crystallize biological macromolecules. (*a*) Two versions of batch crystallization. (*b*) Dialysis method with Cambridge button. (*c*) Vapour-diffusion crystallization with sitting and hanging drops. (*d*) Counter-diffusion setup in X-ray capillary tubes pinched in a gel layer. (*e*) Crystallization by free-interface diffusion in a capillary where two solutions of equivalent volume are brought into contact. The evolution of the macromolecule concentration, $[M]$, and crystallant concentration, $[C]$, in the different methods is indicated (initial and final concentrations in the crystallization solutions are $[M]_i$, $[M]_f$ and $[C]_i$, $[C]_f$, respectively; $[C]_{\rm res}$ is the concentration of the crystallant in the reservoir, and k is a dilution factor specified by the ratio of the initial concentrations of crystallant in the drop and reservoir. In practice, glass vessels in contact with macromolecules should be silicone-treated in a way to obtain hydrophobic surfaces. Typical equilibration trajectories are illustrated in the phase diagram by black arrows, and the evolution of drops after nucleation and during growth are indicated by dashed arrows.

4.1.3.1. *Historical development of methods*

Protein crystallization is an old field that started more than 100 years ago (McPherson, 1991). Early methods included protein extractions with salts or organic solvents, or dialysis of salt solutions against water, and they were carried out on the gramme scale. Batch crystallization was the method of choice at that time. A first breakthrough that paralleled the development of X-ray methods occurred in the 1960s with the development of micromethods such as dialysis and vapour diffusion (with protein at the 1–100 mg scale and crystallization assays in the 10–50 µl range). However, it became rapidly apparent that screening of the entire

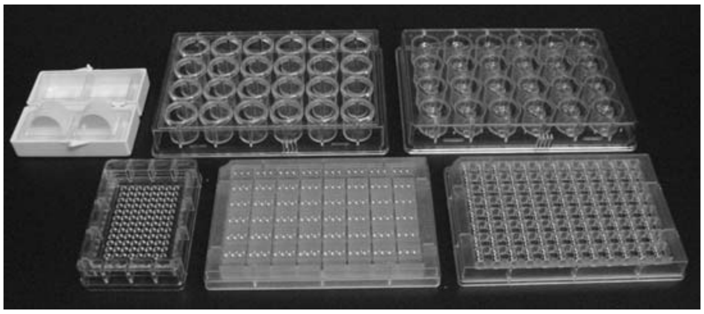


Figure 4.1.3.2
The evolution of crystallization plates from hand-made assays to the high-throughput era. Shown here is an array of crystallization materials and plates. In the upper left is a box of siliconized, glass coverslips used in hanging-drop experiments. Clockwise from the coverslips are a greased VDX plate for hanging-drop experiments, a Cryschem sitting-drop plate, an Intelliplate 96, primarily for use with robotic systems for sitting drops, an Intelliplate 48, and finally, a Vapour–Batch plate that can be used for multiple methods. All of these materials are readily available from commercial sources (photo courtesy of Hampton Research).

crystallization space in searching for appropriate conditions is not feasible in practice. To overcome this difficulty, incomplete factorial methods and other statistical approaches for designing and analysing crystallization experiments were developed (Carter & Carter, 1979; Carter, 1997). A second series of breakthroughs, in the mid-1980s, was stimulated by microgravity research and the advent of structural biology (Giegé *et al.*, 1995). Miniaturization was pursued, molecular biology gave more ready access to samples and more rationality entered the field with, among others, crystallization in gelled media or by counter-diffusion. Moreover, the use of screens was generalized (Jancarik & Kim, 1991) and early robotics facilitated the work of the crystallizers, permitting more extensive screening and yielding improved reproducibility. The last breakthroughs came from structural genomics and have led to further miniaturization (at the nanoscale), more systematic use of robotics and high-throughput screening methods. Finally, advanced methods emerged to produce crystals of better quality (see Section 4.1.4). The size and geometry of crystallization plates recapitulate these methodological developments (Fig. 4.1.3.2).

4.1.3.2. Batch crystallizations

Batch methods are the simplest techniques used to produce protein crystals. They require no more than just mixing the macromolecular solution with crystallants until supersaturation is reached (Fig. 4.1.3.1*a*). Batch crystallization was used to grow crystals in containers of different geometries and samples from the ml to the µl range (McPherson, 1982; Luft *et al.*, 1999*a*). Because one necessarily begins in the supersaturated region of the phase diagram (Fig. 4.1.1.1), nucleation is often excessive. Large crystals, however, can be obtained when the degree of supersaturation is near the metastable region of the crystal/solution phase diagram.

In the early 1990s an automated microbatch crystallization method already allowed the investigation of samples of less than 2 µl (Chayen *et al.*, 1990). In this case, reproducibility is guaranteed. Samples are dispensed and incubated under oil, thus preventing rapid evaporation and uncontrolled concentration changes of the components in the microdroplets. Note, however, that oils are slightly vapour permeable and that the slow evaporation kinetics are dependent on the type of oil (paraffin or silicone oils) used (D'Arcy *et al.*, 1996; Chayen, 1997). The method was adapted for rapid screening of crystallization conditions (D'Arcy, MacSweeney, Stihle & Haber, 2003). It was also adapted for crystallizing proteins in drops suspended between two oil layers (Chayen, 1996; Lorber & Giegé, 1996). Large drops (up to 100 µl) can be deployed allowing easy observation of the crystallization process. Moreover, the absence of contacts between the mother liquor and any solid surfaces results in a lower number of nucleation sites and larger crystals. Note that a hands-free system based on microbatch crystallization in modified X-ray glass capillaries has been developed that allows *in situ* diffraction data collection (Sugahara *et al.*, 2009).

Batch crystallization can also be conducted under high pressure (Lorber *et al.*, 1996) and was adapted for crystallizations on thermal gradients with samples of ~7 µl accommodated in micropipettes (Luft *et al.*, 1999*b*). This latter method permits rapid screening to delineate optimal temperatures for crystallization and frequently yields crystals of sufficient quality for diffraction analysis. Batch methods also permit crystallizations based on thermonucleation (DeMattei & Feigelson, 1993). In a user-friendly version, transfer of crystallization vessels from one thermostated cabinet to another maintained at a higher or lower temperature, depending on whether the protein has normal or retrograde solubility, can readily be done. A variation of classical batch crystallization is the sequential extraction procedure (Jakoby, 1971), based on the property that the solubility of many proteins in highly concentrated salt solutions exhibits significant, but shallow, temperature dependencies.

4.1.3.3. Dialysis methods

Dialysis readily permits easy variation of many parameters that influence protein crystallization. Different types of systems can be used, but all follow the same general principle. The protein is separated from a large volume of solvent by a semi-permeable membrane that allows passage of small molecules, but prevents that of the macromolecules (Fig. 4.1.3.1*b*). Equilibration kinetics depend on the membrane molecular weight exclusion size, the ratio of the concentrations of crystallant inside and outside the macromolecule chamber, the temperature, and the geometry of the dialysis cell. The simplest technique is to use a dialysis bag (*e.g.* of inner diameter ~2 mm), but this usually requires at least 100 µl protein solution per trial.

Crystallization by dialysis has been adapted to small volumes (10 µl or less per assay) in microdialysis cells made from capillary tubes closed by dialysis membranes or polyacrylamide gel plugs (Zeppenzauer, 1971). Microdialysis devices exist in a variety of forms; some are derived from the original Zeppenzauer system (Weber & Goodkin, 1970), others are known as the Cambridge button (Reid *et al.*, 1973) or the microdialysis rod (Lee & Cudney, 2004). With such devices, protein solutions are deposited in 10–50 µl depressions in Plexiglass microdialysis buttons, which are then sealed by dialysis membranes fixed by rubber 'O' rings, and subsequently immersed in a solution contained in the wells of Linbro plates (or other vessels). Wells are sealed with glass coverslips and vacuum grease or tape. In the double-dialysis procedure, the equilibration rate is stringently reduced, thereby improving the method as a means of optimizing crystallization conditions (Thomas *et al.*, 1989). Equilibration rates can be manipulated by choosing appropriate membrane molecular weight exclusion limits, distances between dialysis membranes or relative volumes.

4.1.3.4. Vapour-diffusion methods

Crystallization by vapour diffusion was introduced in structural biology for the preparation of tRNA crystals (Hampel *et al.*, 1968). It is well suited for small volumes (as little as 1 µl or less) and has become the favoured method of many experimenters. It is practised in a variety of forms and is the method of choice for robotic applications. In all of its versions, a drop with the protein to be crystallized, together with buffer, crystallant and additives, is equilibrated against a reservoir containing a solution of crystallant at a higher concentration than that in the drop (Fig. 4.1.3.1*c*). Equilibration proceeds by diffusion of the volatile species until the vapour pressure of the drop equals that of the reservoir. If equilibration occurs by water (or organic solvent) exchange from the drop to the reservoir (*e.g.* initial salt concentration in the reservoir higher than in the drop), it leads to a volume decrease of the drop, so that the concentration of all constituents in the drop increases. The situation is reversed if the initial concentration of the crystallizing agent in the reservoir is lower than that in the drop. In that case, water exchange occurs from reservoir to drop. Crystallization of several macromolecules was achieved using this 'reversed' procedure (Giegé *et al.*, 1977; Richard *et al.*, 1995; Jeruzalmi & Steitz, 1997).

Historically, hanging drops have been deployed in Linbro-type tissue culture plates (Fig. 4.1.3.2). These plates contain 24 wells with volumes of ~2 ml and each well is covered by a glass coverslip sealed with vacuum grease. Drops are composed by mixing 2–10 µl aliquots of the protein with aliquots of the crystallant and additional components, as needed. A variant of the procedure is the HANGMAN method. It utilizes a clear, non-wetting adhesive tape that both supports the protein drops and seals the reservoirs (Luft & De Titta, 1992). Modern arrangements either maintain the 24-well format with more convenient screw-in lids, or adopt the 96-well format with drops deposited on an adhesive sealing tape. The general principle remains the same: a twofold ratio between the concentration of the crystallant in the reservoir and in the drop is most frequently used. This is achieved by mixing a droplet of protein at twice the desired final concentration with an equal volume of the reservoir at the proper concentration (to prevent drops from falling into the reservoir, their final volume should not exceed 15 µl). When no crystals or precipitate are observed in the drops, either sufficient supersaturation has not been reached or only the metastable region has been attained. In the latter case, changing the temperature by a few degrees may be sufficient to initiate nucleation. In the former case, the concentration of precipitant in the reservoir must be increased.

Sitting drops can be dispensed into a variety of different devices. Arrangements consisting of Pyrex plates with a variable number of depressions (up to nine) installed in sealed boxes were used for tRNA crystallization (Dock *et al.*, 1984). Drops of mother liquor are dispensed in the depressions and reservoir solutions with crystallant are poured into the bottom sections of the boxes. These systems are efficient for large drop arrays and can be used for both screening and optimizing crystallization conditions. Multichamber arrangements are suitable for the control of individual assays (Fig. 4.1.3.2). They often consist of polystyrene plates with a variable number of wells (24, 96 and more) that can be individually or collectively sealed. Sitting drops can also be placed on microbridges (Harlos, 1992) or supported by plastic posts in the centres of the wells. Reservoir solutions are contained in the wells in which the microbridges or support posts are placed. Plates with 96 wells, sealed with clear sealing tape, are convenient for large matrix screening. Most of these plates are commercially available and can often be used for a majority of different vapour-diffusion crystallization methodologies (hanging, sitting or sandwich drops, the latter being maintained between two glass plates). A crystallization setup in which drops are deployed in glass tubes, which are maintained vertical and epoxy-sealed on glass coverslips, is known as the plug-drop design (Strickland *et al.*, 1995). Plug-drop units are placed in the wells of Linbro plates surrounded by reservoir solution and then the wells are sealed as usual. With this geometry, crystals do not adhere to glass coverslips as they may with sandwich drops.

Vapour-phase equilibration can be achieved in capillaries (Luft & Cody, 1989) or even directly in X-ray capillaries, as was described for ribosome crystallization (Yonath *et al.*, 1982). This last method may even be essential for fragile crystals, where transferring from crystallization cells to X-ray capillaries or cryo-loops can produce mechanical damage. Vapour-diffusion methods permit easy variations of physical parameters during crystallization, and many successes have been obtained by affecting supersaturation by temperature or pH changes. With ammonium sulfate as the crystallant, it has been shown that the ultimate pH in the drops is imposed by the pH of the reservoir (Mikol *et al.*, 1989). Thus, varying the pH of the reservoir permits adjustment of that in the drops. Sitting drops are also well suited for carrying out epitaxic growth of macromolecule crystals on mineral matrices or other surfaces (McPherson & Shlichta, 1988*a*; Kimble *et al.*, 1998; Stolyarova *et al.*, 2006).

The kinetics of water evaporation (or of any other volatile species) determine the kinetics of supersaturation and, consequently, that of nucleation (Martins, Pëssoa *et al.*, 2008; Martins, Rocha & Damas, 2008). Kinetics measured from hanging drops containing ammonium sulfate, PEG or MPD are significantly influenced by experimental conditions (Mikol, Rodeau & Giegé, 1990; Luft *et al.*, 1996). The parameters that chiefly determine equilibration rates are temperature, initial drop volume (and initial surface-to-volume ratio of the drop and its dilution with respect to the reservoir), water pressure and the chemical nature of the crystallant (Newman, 2005; Newman *et al.*, 2007). The distance separating the drop from the reservoir solution is another critical element and a simple procedure was designed to allow variation of the rate of water equilibration, thereby optimizing crystal-growth conditions (Luft *et al.*, 1996). Evaporation rates can also be monitored and controlled in a weight-sensitive device (Shu *et al.*, 1998). Another method uses oil layered over the reservoir, which permits very slow evaporation of the underlying aqueous solution (Chayen, 1997). The thickness of the oil layer, therefore, dictates evaporation rates and, consequently, crystallization rates (see Section 4.1.3.2).

The period for water equilibration to reach 90% completion can vary from ~25 h to more than 25 d. The most rapid equilibration occurs with ammonium sulfate, more slowly with MPD, while that with PEG is by far the slowest. An empirical model has been proposed which estimates the minimum duration of equilibration under standard experimental conditions (Mikol, Rodeau & Giegé, 1990). Equilibration, which brings the macromolecules very slowly to a supersaturated state, may explain the crystallization successes with PEG as the crystallant. This explanation is corroborated by experiments showing an increase of the terminal crystal size when equilibration rates are reduced (Chayen, 1997).

4.1.3.5. Free-interface and counter-diffusion methods

In these methods, equilibration occurs by direct diffusion of the crystallant into the protein solution (Salemme, 1972; Garcia-Ruiz & Moreno, 1994). Both methods require minimal convection and, therefore, experiments are conducted in capillaries (Fig. 4.1.3.1*d*,*e*). To avoid too rapid mixing in a free-interface diffusion assay, the less dense solution is poured gently on the densest one. One can also freeze the solution with the crystallant and layer the protein solution on top. The components of each solution in contact will then diffuse into the other compartment.

A more versatile version of this technique is counter-diffusion crystallization (Garcia-Ruiz, 2003; Ng *et al.*, 2003; Otálora *et al.*, 2009), which was referred to as the 'gel acupuncture' method in its first version (Garcia-Ruiz & Moreno, 1994). In a typical experiment, a gel base is formed from agarose or silica in a small container, and a concentrated crystallant solution is poured over its surface. This crystallizing agent permeates the gel by diffusion, forming a gradient. A capillary filled with the protein is inserted into the gel. The crystallant then enters the capillary from the gel and forms an upward gradient in the capillary, promoting crystallization along its length as it rises by pure diffusion. The method operates with a variety of gels and crystallants, with different heights of these agents over the gel and with open or sealed capillaries. It has already been successfully used to crystallize a variety of macromolecules, some of very large size (Garcia-Ruiz *et al.*, 1998; Biertümpfel *et al.*, 2002; Kutá Smatanová *et al.*, 2006).

The main advantage of the counter-diffusion technique is that a wide range of supersaturation conditions is tested in a single experiment. While a precipitate may form at the entrance of the capillary at high supersaturation, monocrystals may grow at the opposite end where supersaturation is lower. A second advantage is that all steps from crystallization to structure solution, including substrate/cryoprotectant/heavy atoms soaking and X-ray analysis can be performed without any crystal handling (Gavira *et al.*, 2002). In recent versions, the diameter and length of counter-diffusion capillaries have been reduced so that assays can be conducted at a 300 nl scale (see Section 4.1.4.1.).

4.1.3.6. Miniaturization, automation and robotics

The first user-friendly attempt to miniaturize and automate crystallization concerned batch crystallization in microdroplets under oil (Chayen *et al.*, 1990). In recent years, robots and other automated instruments, and entire integrated systems, have been developed to accelerate the crystallization process (DeLucas *et al.*, 2003; Hosfield *et al.*, 2003; Luft *et al.*, 2003; Bard *et al.*, 2004; Berry *et al.*, 2006) and the optimization procedures (Newman *et al.*, 2008). They have the capacity to screen thousands of crystallization conditions, and they do so precisely and reliably, with fewer errors and better record keeping than most humans. In many large laboratories, these have become essential pieces of equipment. Using standard, usually commercially available, screening kits, sometimes supplemented by local preferences or family-directed screens, they can often arrive at acceptable crystals in the most expeditious manner possible.

Robotic systems are efficient, tireless and accurate, but they offer another important feature in addition. They can carry out experiments using drop samples of very small volume, drops of 1 µl in most cases, nanolitres in some. This, in turn, produces a requirement for automated, microphotographic visualization instruments, and complex storage and handling systems, and their associated expense. On the other hand, a great advantage emerges in that they can perform enormous numbers of crystallization trials using remarkably little biological sample. This, in turn, relieves the investigator of a significant burden in terms of preparing and purifying macromolecules.

Many of the robotic systems are based on reproducing procedures currently used for manual experiments, such as sitting and hanging drops, and microdrops under oil. They are simply carried out on a much smaller scale. More recently, however, even more miniaturized devices have come on the market. These use what is now commonly called nanotechnology to manipulate small amounts of liquids and fluid streams. These devices are only now seeing rigorous evaluation in laboratories, but they clearly show great promise for the future (see Section 4.1.4). Another effort is underway to develop robotic systems for crystal harvesting and crystallization devices that will allow direct X-ray exposure of crystals where they are grown *in situ* (Viola *et al.*, 2007). These would obviate the need for careful mounting, an often-problematic aspect of data collection.

4.1.4. Advanced crystallization methodologies

In methods that manipulate physical parameters, the effects on crystallization are manifold. Among others, they may influence fluid properties in the crystallization media and movement of molecules (gravity), alter the conformation of the macromolecule (pressure), orient crystals (magnetic field), or influence nucleation (electric field). Thus, initiation of crystallization may be triggered by various mechanisms, growth may be differently influenced and, in favourable cases, crystal quality improved.

4.1.4.1. Crystallization in convection-free media

Theoretical considerations. When a crystal starts to grow, it attracts surrounding molecules and creates a concentration gradient. Since crystallization occurs on earth in the gravity field, this gradient of concentration and density will lead to convective currents in the mother liquor. In addition, as soon as the crystal becomes big enough, it will sink to the bottom of the solution. Convection and sedimentation almost always take place in classical experimental setups and they almost certainly influence crystallization processes. Their contribution would be drastically reduced in the absence of gravity, as occurs in weightlessness, and the theory predicts more regular crystal growth under a microgravity–diffusive regime that should favour enhanced crystal quality. Such considerations have justified space-crystallization programmes and, as a consequence, have contributed to a deeper understanding of the crystallization process of biomacromolecules (Giegé *et al.*, 1995; McPherson, 1996; Kundrot *et al.*, 2001). However, because of limited access to space experimentation, crystallization in weightlessness will never be user friendly. This has stimulated studies for finding easy ways to simulate microgravity conditions in the laboratory. Such methods, where crystal growth is less dependent on convection and more on diffusion, take advantage of gelled media and microfluidic environments.

Use of microgravity. The first observation in microgravity was that the absence of sedimentation permits the growth of individual crystals in suspension, without any perturbation by contact with vessel walls and neighbouring crystals. However, one should bear in mind that even in microgravity small accelerations can occur owing to vehicle movement and crystal displacement has been recorded (*e.g.* Lorber *et al.*, 2000). Microgravity experiments require specific instrumentation with dedicated reactors based on current batch, dialysis and vapour-diffusion methods, or on more microgravity-relevant approaches, such as free-interface or counter-diffusion, with crystallization vessels often of rather large size (DeLucas *et al.*, 1994; Giegé *et al.*, 1995; McPherson, 1996; Gonzalez-Ramirez, Carrera *et al.*, 2008). Reproducible data, with a substantial number of model proteins (lysozyme, thaumatin, canavalin and several plant viruses), were obtained in the versatile Advanced Protein Crystallization Facility instrument from the European Space Agency (Vergara *et al.*, 2003, 2005). Altogther, an overall positive effect of microgravity on protein crystal growth emerged.

In support of this conclusion are observations of larger sizes for space-grown crystals and improved optical quality, as exemplified by thaumatin crystals (Ng, Lorber *et al.*, 1997). The maximum resolution of diffraction patterns also indicated superiority for microgravity crystals. A striking example is parvalbumin that diffracts to 0.9 Å resolution, while the earth-grown crystals are not suitable for diffraction analysis (Declercq *et al.*, 1999). Also, in several instances, the signal-to-noise ratio of X-ray diffraction data collected from space crystals was greater than for the corresponding earth controls, as for satellite tobacco mosaic virus (McPherson, 1996) and thaumatin (Ng, Lorber *et al.*, 1997; Lorber, Sauter, Robert *et al.*, 1999). An additional criterion is the reduced mosaic spread of reflections recorded from space samples (Snell *et al.*, 1995; Stojanoff *et al.*, 1996; Ng, Lorber *et al.*, 1997; Lorber, Sauter, Robert *et al.*, 1999). Impurity incorporation during growth is another issue and, as shown with lysozyme, the microgravity-grown crystals incorporate fewer impurities than the earth controls (Carter *et al.*, 1999). The best criterion for enhanced crystal quality, however, is the crystallographic structure. In a case study with lysozyme, significant improvement of resolution from 1.6 to 1.35 Å, decreased atomic displacement parameters (ADPs or B factors) and a structure of increased clarity have been noted for the space-grown crystals (Carter *et al.*, 1999). In another study on an aspartyl-tRNA synthetase, a strictly comparative analysis showed that crystals grown in microgravity were superior in many respects to controls prepared under otherwise identical conditions on earth, facilitating structure determination at 2.0 Å resolution (Ng *et al.*, 2002).

Nevertheless divergent conclusions were reached concerning the quality of the X-ray structure, which was shown to be improved (Carter *et al.*, 1999) or unaffected (Vaney *et al.*, 1996) by microgravity. This contradiction may originate from different levels of impurities present in the protein samples and/or from non-identical growth conditions in different hardware. Conceivably, alteration by gravity of fluid properties could affect nucleation. Transport is of importance, because the large sizes of proteins imply that they have low diffusivities. Elimination of fluid convection may, however, dramatically affect the movement and distribution of proteins in the fluid and their transport and absorption to crystal surfaces. In addition, many proteins form non-specific aggregates in solution. These may be a major source of the contaminants that are incorporated into crystal lattices. By virtue of their size and low diffusivity, the movement of aggregates and large impurities in solution is even more significantly altered.

On earth there is continuous density-driven convective mixing in the solution due to gradients arising from temperature or from incorporation of molecules by the growing crystal. The effects of diffusive transport in the laboratory are, by comparison with the microgravity case, almost negligible because of the very slow rate of diffusion of large proteins. Because of convective mixing, protein crystals nucleated on earth are continuously exposed to the full concentration of protein nutrient present in the bulk solvent. Convection thus maintains, at the growing crystal interface, excessive supersaturation as growth proceeds. This provides an explanation as to why microgravity may improve the quality of protein crystals. The mechanism for enhanced order and reduction of defects may not be directly due to convective turbulence at growing crystal surfaces, but to reduction of the concentration of nutrient molecules and impurities in the immediate neighbourhood of the growing crystals. As a protein crystal forms in microgravity, a concentration gradient or 'depletion zone' is established around the nucleus. Because protein diffusion is slow and that of impurities may be even slower, the depletion zone is quasi-stable. The net effect is that the surfaces of the growing crystal interface with a local solution phase at a lower concentration of protein nutrient and impurities than exists in the bulk solvent. The crystal, as it grows, experiences a reduction in its local degree of supersaturation and essentially creates for itself an environment equivalent to the metastable region where optimal growth is expected.

Investigations of protein crystallization diverged along two paths. The objective of the first was to produce high-quality crystals for X-ray diffraction analysis. The crystals themselves were the product of the space-bound experiment and biochemical results were secondary. The goal of the second line of investigation was to understand and to control the physics of the process. This second interest was supported by extensive ground-based research. The confluence of results yielded persuasive explanations for the observed improvements in size and quality of protein crystals grown in microgravity and a robust theoretical framework for understanding the phenomena involved. They showed also that protein crystals are more sensitive to the very high degrees of supersaturation at which they are usually grown and to the mass-transport mechanisms responsible for bringing nutrient to their growing surfaces. The self-regulating nature of protein crystallization in microgravity, through the establishment of local concentration gradients of reduced supersaturation, explains why the diffusive transport that predominates produces a significant difference in ultimate crystal quality.

Crystallization in gelled media. Because convection occurs in free solutions, crystallization in gels represents what is essentially a convection-free environment (Henisch, 1988). Thus, the quality of crystals may be improved in gels. Whatever the mechanism of crystallization in gels, the procedure will produce changes in the nucleation and crystal-growth processes, as has been verified with many proteins (Robert & Lefaucheux, 1988; Cudney *et al.*, 1994; Vidal *et al.*, 1999; Biertümpfel *et al.*, 2002; Lorber *et al.*, 2009). Two types of gels have been used, namely, agarose and silica gels. The latter seem to have proven the most adaptable, versatile and useful for proteins. With both agarose and silica gels, it is possible to use a variety of different crystallants, including salts, organic solvents and polymers such as PEG (Gonzalez-Ramirez, Caballero & Garcia-Ruiz, 2008). They also allow the investigator to control pH and temperature. The most successful efforts have involved direct diffusion arrangements, where the crystallant is diffused into a protein-containing gel or *vice versa*. In practice, experiments are conducted in semi-liquid gelled media where the agarose concentration does not exceed 0.6%(w/v), but crystallization can also take place in 2.0%(w/v) agarose viscoelastic gels, a condition that does not affect the crystal structures (Sugiyama *et al.*, 2009).

Crystals grown in gels can easily be removed from their soft environment and set up for X-ray analysis. They tend to be robust since, as shown with lysozyme crystals, agarose fibres are incorporated into the crystal lattice (Gavira & Garcia-Ruiz, 2002). Gel

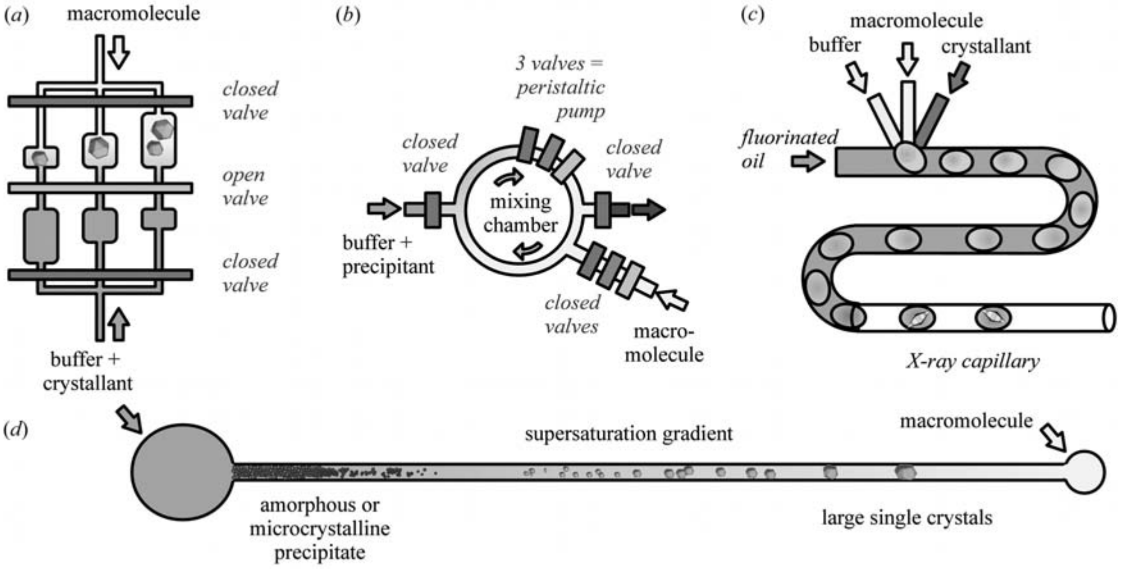


Figure 4.1.4.1
Examples of microfluidic devices designed for biocrystallization. (*a*) Schematic view of one of the 46 modules composing the free-interface diffusion chip. An integrated fluidic circuit dispenses the macromolecule and crystallant solutions into the final chambers. The microchannels are then closed and those connecting the top and bottom chambers are opened with the help of pneumatic valves integrated in the chip. The crystallant diffuses in the protein chamber and triggers crystallization. In this version of the chip, each module is divided into three pairs of chambers with volumes of 5–20 nl to create different protein-to-crystallant concentration ratios. (*b*) The mixing rotor of a formulation chip designed for the high-throughput study of precipitation diagrams. This chip generates protein/buffer/crystallant mixtures at different concentrations in its 5 nl rotor. The three valves in a row constitute a peristaltic pump that homogenizes the mixture. A charge-coupled device (CCD) camera is used to detect the appearance of a precipitate. (*c*) The nanobatch chip. Nanodroplets are produced in a microfluidic channel and displaced by inert oil (the flow rate determines the drop size from 10 to 20 nl). Droplets can be stored in the chip or in capillaries connected to the exit of the chip. They can be inspected and crystals can be characterized by X-ray diffraction. (*d*) Microfluidic chip for counter-diffusion experiments. This method relies on the diffusion of a crystallant into an elongated chamber (the microfluidic channel) containing a macromolecular solution. A concentration gradient is generated that develops along the entire crystallization chamber. The propagating supersaturation wave of gradually decreasing amplitude tests a broad range of nucleation and growth conditions in a single experiment. While a precipitate may form at the entrance of the chamber, monocrystals may grow at the opposite end. Crystals can be observed and analysed by X-ray diffraction directly inside the chip (adapted from Sauter *et al.*, 2007).

growth, because it suppresses convection, has proven to be a useful technique for analysing concentration gradients around growing crystals by interferometric techniques (Robert *et al.*, 1994) and growth mechanisms by differential interference contrast microscopy (Van Driessche, Otalora, Gavira & Sazaki, 2008). In conclusion, gelled media mimic microgravity conditions, preserve crystals once they are grown and, as expected, crystals grown in gels are often of superior quality to controls grown from solutions (Zhu *et al.*, 2001; Moreno *et al.*, 2002; Sauter *et al.*, 2002). Finally, gels can prevent damage during crystal soaking and cryo-cooling (Biertümpfel *et al.*, 2005; Sauter *et al.*, 2009; Lorber *et al.*, 2009).

Crystallization in microfluidic devices. Microfluidic devices were recently introduced in the field of biological crystal growth and represent a new means of crystallizing under diffusive conditions. These systems were primarily intended to miniaturize and to parallelize crystallization assays, thus leading to novel, cost-effective, high-throughput screening approaches. However, because of the small size of their channels and chambers (typically below 100 µm in depth and width), they also provide a diffusive environment comparable to that existing in a capillary tube, in a gel, or under microgravity.

Indeed, the first microfluidic application in biocrystallization was a miniaturized version of the free-interface diffusion technique in which the absence of convection is essential (Hansen *et al.*, 2002). The chip consists of a complex integrated fluidic circuit including two networks of channels, one for liquid handling and a second serving as actuation valves. The chip is dedicated to high-throughput screening and, in its initial version, was designed to test 48 crystallization conditions with less than 10 µl of sample solution. Three parallel sets of chambers are used to bring into contact different proportions of macromolecule and crystallant solutions (Fig. 4.1.4.1*a*). This concept of chip was further modified to combine free-interface diffusion with vapour diffusion (or vapour permeation) for fine tuning the supersaturation achieved in crystallization chambers (Hansen *et al.*, 2006). This technology also led to a 'formulator chip' that can perform hundreds of mixing operations in just a few hours in order to establish precipitation diagrams (Fig. 4.1.4.1*b*). A single assay consumes less than 5 nl sample/buffer/precipitant solution and derived precipitation maps are used to delineate a grid of conditions for crystallization screening (Hansen *et al.*, 2004).

The second crystallization method implemented in microfluidics was 'batch in nanodroplets' (Zheng *et al.*, 2003). The chip design is extremely simple: it consists of inlets for protein, buffer and crystallant solutions, and a microfluidic channel in which 10 nl droplets are prepared by mixing these solutions in various ratios. This device allows a daily formulation of thousands of nanodrops or plugs (Fig. 4.1.4.1*c*), which are carried by a flow of inert fluorocarbon oil. They are stored on the chip or in capillary tubes plugged at the exit of the chip and their content can easily be analysed by X-ray diffraction (Yadav *et al.*, 2005). This method is very well suited for high-throughput screening and, in addition to crystal growth, this technology can be used for many applications in chemistry (Song *et al.*, 2006).

Based on the nanodrop approach, a more complex system has recently been designed for basic research purposes. It is able to formulate droplets and to flow them to storage chambers where they can be concentrated or diluted by water permeation through the chambers' walls. This 'phase chip' is designed to establish phase diagrams with total control over supersaturation, nucleation and growth kinetics in each individual drop (Shim *et al.*, 2007).

As for free-interface diffusion, the absence of convection in microfluidic channels makes microsystems very appealing for implementing counter-diffusion experiments (Sauter *et al.*, 2007; Ng *et al.*, 2008). Characteristic counter-diffusion features were successfully reproduced in microchannels with a production of crystalline material ranging from microcrystals to large monocrystals along the supersaturation gradient. When made of appropriate polymer material, these counter-diffusion chips also allow a direct on-chip characterization of the crystals by X-ray diffraction, without any further (and potentially deleterious) sample handling (Ng *et al.*, 2008; Dhouib *et al.*, 2009).

These examples illustrate the many advantages of microfluidic chips – low sample consumption, high-throughput screening capabilities, quasi-ideal convectionless growth conditions – and one can anticipate that microfluidic technology will become a popular and affordable tool both for condition screening, optimization and X-ray analysis, and for basic crystallogenesis research.

Simulating other aspects of microgravity crystal growth. Heterogeneous nucleation or crystal growth on the solid surface of crystallization vessels can be avoided under levitation (Rhim & Chung, 1990) and more easily in batch between two oil layers (Chayen, 1996; Lorber & Giegé, 1996). This can also be achieved for the growth of large protein crystals by mild stirring of the solution in two-liquid systems (Adachi *et al.*, 2004).

It was conjectured that other features of weightlessness, such as suppression of convection, could be achieved in the laboratory under hypergravity and when magnetic or electric fields are applied. These possibilities have been tested experimentally. Crystals were grown under forced diffusive transport of the macromolecules in centrifuges (Karpukhina *et al.*, 1975; Lenhoff *et al.*, 1997; Lorber, 2008) and nucleation was shown to be affected by magnetic (Ataka *et al.*, 1997; Sazaki *et al.*, 1997) or external electric (Taleb *et al.*, 1999; Nanev & Penkova, 2001) fields. Interestingly, under these last conditions, growing crystals were shown to have preferential orientations and specific spatial distributions in the crystallization chambers. Magnetic fields produced by small permanent magnets of 1.25 T are sufficient to produce these effects (Astier *et al.*, 1998) and numerical predictions revealed that magnetization forces could damp convection (Qi *et al.*, 2001). For crystallization induced by electric fields, simple devices adapted to vapour-diffusion (Charron *et al.*, 2003) and batch (Al-Haq *et al.*, 2007) methods are available. Crystallization can also be electrochemically assisted by internal electric fields (Frontana-Uribe & Moreno, 2008). In some cases, magnetic and electric fields have been coupled and experiments conducted in gelled media (Sazaki *et al.*, 2004; Moreno *et al.*, 2009).

Although the above methods are not widespread and the underlying physics not completely validated, they can be useful in special cases. For instance, when crystallization attempts systematically yield showers of microcrystals, crystallization inside electric or magnetic fields can be an alternative to obtain monocrystals suitable for X-ray data collection, because the number of nucleation sites is reduced and can be controlled (Moreno & Sazaki, 2004; Hammadi *et al.*, 2009) and crystal quality maintained (Sato *et al.*, 2000; Lübbert *et al.*, 2004).

4.1.4.2. Methods making use of temperature and pressure

Temperature and pressure are familiar thermodynamic parameters. Indeed, many living organisms are thermophiles, even hyperthermophiles, or barophiles/piezophiles and thus have evolved macromolecules stable at temperatures up to 110 °C or pressures up to 100 MPa, *i.e.* 1000-fold atmospheric pressure (Abe & Horikoshi, 2001). Temperature can trigger nucleation, regardless of the crystallization method. This can be done in a controlled manner, but often occurs as an unexpected consequence of accidental temperature variation in the laboratory. Dedicated systems have been designed for temperature-dependent control of nucleation and growth (Astier & Veesler, 2008), and find application for, among other things, the growth of large high-quality protein crystals for neutron crystallography (Budayova-Spano *et al.*, 2007).

Pressure, as anticipated, can trigger nucleation and sustain protein crystal growth (*e.g.* Suzuki *et al.*, 2002). To facilitate analysis of crystallization output, assays under pressure can be done in agarose gel (Kadri *et al.*, 2003). Rather simple equipment is required allowing batch crystallization of ~12 individual samples of ~80 µl that can be collectively pressurized up to ~400 MPa (Lorber *et al.*, 1996). The effects exerted by pressure are multiple and protein dependent, with habit, number, length, shape and solubility of crystals modified under pressure. Further, crystallization volumes and diffraction properties are affected and, interestingly, these physical properties are essentially conserved upon depressurization of the crystals. In particular, differences in the water sites surrounding thaumatin crystals grown at 0.1 and 150 MPa have been observed (Charron *et al.*, 2002). Crystallographic analysis of cowpea mosaic virus crystals compressed at 330 MPa in a diamond-anvil cell demonstrated pressure-induced ordering of the crystals, lower ADPs and a larger number of ordered water molecules (Girard *et al.*, 2005; Lin *et al.*, 2005).

4.1.4.3. Methods making use of crystallization chaperones

Another strategy that has been used for recalcitrant proteins is to combine them in some manner with a second protein, sometimes called a cocrystallization or chaperone protein (Warke & Momany, 2007; Koide, 2009), so that the complex of the two provides an additional chance for success. The idea was first tested with lysozyme complexed with an Fab antibody fragment (Boulot *et al.*, 1988) and has been used particularly with membrane proteins and antibody domains directed against the target protein (Ostermeier *et al.*, 1995). In those cases the antibody fragment enhanced the solubility of the otherwise hydrophobic protein and provided additional lattice contacts in the resultant crystals. There is, in principle, no reason why such 'crystallization chaperones' could not be used with soluble proteins. Likewise the method can be useful for the crystallization of functional RNA fragments (Ye *et al.*, 2008). An alternate possibility with great potential is the recently developed DARPin technology based on the natural ankyrin repeat protein fold with randomized surface residue positions allowing specific binding to virtually any target protein (Sennhauser & Grütter, 2008).

4.1.4.4. Seeding

It is often necessary to reproduce crystals grown previously, where either the formation of nuclei is limiting, or spontaneous nucleation occurs at such a profound level of supersaturation that poor growth results. In such cases, it is desirable to induce growth in a directed fashion at low levels of supersaturation. This can be accomplished by seeding a metastable, supersaturated protein solution with crystals from earlier trials. Seeding also permits one to uncouple nucleation and growth. Seeding techniques fall into two categories that employ either macroseeds (Thaller *et al.*, 1985) or microcrystals as seeds (Stura *et al.*, 1999; Bergfors, 2003). In both cases, the solution to be seeded should be only slightly supersaturated so that controlled growth can occur.

When seeding with crystals large enough to be manipulated under a microscope, the most important consideration is to eliminate spurious nucleation by transfer of too many seeds. This drawback may be overcome using laser tweezers, a technique that permits non-mechanical, *in situ* manipulation of individual seeds as small as 1 µm (Bancel *et al.*, 1998). Even if a single large crystal is employed, microcrystals adhering to its surface may be carried across to the fresh solution. To avoid this, the macroseed is washed by passing it through a series of intermediate transfer solutions. In doing so, not only are microcrystals removed, but if the wash solutions are chosen properly, some limited dissolution of the seed surface may take place. This has the effect of freshening the seed-crystal surfaces and promoting new growth once it is introduced into the new protein solution. Note that crystals of homologous macromolecules can serve as seeds (Thaller *et al.*, 1985).

In the second approach with microcrystals, the danger is that too many nuclei will be introduced into the fresh supersaturated solution, and masses of crystals will result. To overcome this, a stock solution of microcrystals is serially diluted over a very broad range. Some dilution sample in the series will, on average, have no more than one microseed per ml; others will have several times more, or none at all. An aliquot (~1 µl) of each sample in the series is then added to fresh crystallization trials. This empirical test, ideally, identifies the correct sample to use for seeding by yielding only one or a small number of single crystals when crystal growth is completed. Microseeds can be introduced into crystallization trials at any stage of microbatch or vapour-diffusion experiments (D'Arcy, MacSweeney & Haber, 2003; D'Arcy *et al.*, 2004) and this process can be automated (D'Arcy *et al.*, 2007; Newman *et al.*, 2008; Khurshid *et al.*, 2010).

4.1.5. From the macromolecule to perfect crystals: the physics view

Each of the four stages in crystallization (prenucleation, nucleation, growth and cessation of growth) can be monitored by specific physical techniques. Although systematic characterization of crystallization is usually not carried out in practice, characterization of individual steps and measurement of the physical properties of crystals obtained under various conditions may help in the design of appropriate experimental conditions to reproducibly obtain crystals of a desired quality (*e.g.* of larger size, improved morphology, increased resolution or greater perfection).

4.1.5.1. Prenucleation and nucleation

DLS relies on the scattering of monochromatic light by aggregates or particles moving in solution. Since diffusivity of the particles is related to their size by the Stokes–Einstein equation, measurement of diffusion coefficients can be translated into hydrodynamic radii. By making measurements as a function of scattering angle, information regarding aggregate shape can also be obtained. For single-component systems, the method for determining the size of macromolecules, viruses and larger particles up to a few µm is straightforward. For polydisperse and concentrated systems, the problem is more complex, but with the use of auto-correlation functions and advances in signal detection, DLS provides good estimates of aggregate size distribution and supplies a diagnostic tool for quality control of proteins and optimization of crystallization conditions. Experimental necessity encouraged the design of dedicated instruments for protein crystallization, combining *e.g.* imaging of crystals and DLS analysis within crystallization droplets (Dierks *et al.*, 2008).

In biocrystallogenesis, investigations based on light scattering have been useful in detecting nucleation prior to the appearance of crystals observable under the light microscope, that is, in understanding prenucleation and nucleation. Many studies have been carried out with lysozyme as the model (Kam *et al.*, 1978; Durbin & Feher, 1996), though not exclusively, and they have been developed with two objectives. One is to analyse the kinetics and the distribution of molecular aggregate sizes as a function of supersaturation. The idea is to understand the nature of prenuclear clusters that form in solution and how they transform into crystal nuclei (Kam *et al.*, 1978; Georgalis *et al.*, 1993; Malkin & McPherson, 1994). Such a quantitative approach has sought to define the underlying kinetic and thermodynamic parameters that govern the nucleation process.

A more practical objective is to use light-scattering methods to predict which combinations of crystallants, additives and physical parameters are most likely to lead to the nucleation and growth of crystals (Mikol, Hirsch & Giegé, 1990; Ferré-D'Amaré & Burley, 1997; Borgstahl, 2007; Wilson, 2003; Niesen *et al.*, 2008). A major goal here is to reduce the number of empirical trials. The analyses depend on the likelihood that precipitates are usually linear, branched and extended in shape, since they represent a kind of random polymerization process (Kam *et al.*, 1978). Aggregates leading to nuclei, on the other hand, tend to be more globular and three dimensional in form. Thus, mother liquors that indicate a nascent precipitate can be identified as a failure, while those that have the character of globular aggregates hold promise for further exploration and refinement. Other analyses have been based on discrimination between polydisperse and monodisperse protein solutions, which suggests that polydispersity hampers crystallization, while monodispersity favours it (Mikol, Hirsch & Giegé, 1990).

A more quantitative approach is based on measurement of the second virial coefficient B2, which serves as a predictor of the type of interaction between macromolecules occurring in solution. Using static light scattering, it was found that mother liquors that invariably yield crystals have second virial coefficients that fall within a narrow range of small negative values. Correlations between the associative properties of proteins in solution, their solubility and the B2 coefficient were highlighted (*e.g.* George *et al.*, 1997; Wilson, 2003), and seem to be a general feature. This is a powerful diagnostic of crystallization conditions.

Related methods, such as fluorescence spectroscopy (Crosio & Jullien, 1992; Forsythe *et al.*, 2006), osmotic pressure (Bonneté *et al.*, 1997; Neal *et al.*, 1999), small-angle X-ray scattering (Finet *et al.*, 1998) and small-angle neutron scattering (Ebel *et al.*, 1999; Gripon *et al.*, 1997; Minezaki *et al.*, 1996; Vidal *et al.*, 1998), were used to investigate specific aspects of protein interactions under precrystallization conditions and produced, in several instances, complementary answers to those from light-scattering studies.

4.1.5.2. Growth and cessation of growth

A number of microscopies and other optical methods can be used for studying the crystal growth of proteins (Van Driessche, Otalora, Sazaki *et al.*, 2008). These are time-lapse video microscopy with polarized light, Schlieren and phase-contrast microscopy, Mach–Zehnder and phase-shift Mach–Zehnder interferometry, Michelson interferometry, electron microscopy (EM), atomic force microscopy (AFM) and laser confocal microscopy with differential interference contrast microscopy. Each of these methods provides complementary data which, in combination, have yielded answers to many relevant questions.

Time-lapse video microscopy has been used to measure growth rates (Koszelak & McPherson, 1988; Lorber & Giegé, 1992; Zhu *et al.*, 2001). It was valuable in revealing unexpected phenomena, such as capture and incorporation of microcrystals by larger crystals, contact effects, consequences of sedimentation, flexibility of thin crystals, fluctuations in growth rates and initiation of twinning (Koszelak *et al.*, 1991). Optical microscopy and interferometric methods gave information on concentration gradients that appear as a consequence of incorporation of molecules into the solid state. These methods, however, suffer from rather shallow response dependence with respect to protein concentration. This can be overcome by the introduction of phase-shift methods, as has been successfully achieved in the case of Mach–Zehnder interferometry. With this technique, gradients of protein concentration have been mapped in the mother liquor and around growing crystals. Classical Mach–Zehnder interferometry has been used to monitor diffusion kinetics and supersaturation levels during crystallization, as was done in dialysis setups (Snell *et al.*, 1996) or in counter-diffusion crystal growth cells (Garcia-Ruiz *et al.*, 1999). Using laser confocal microscopy combined with differential interference contrast microscopy it was possible to visualize dislocations in protein crystals during growth (Sazaki *et al.*, 2005). Single-molecule visualization techniques gave access to direct observation of the diffusion of individual fluorescence-labelled protein molecules at an interface between a solution and a protein crystal (Sazaki *et al.*, 2008).

Michelson interferometry can be used for direct growth measurements on crystal surfaces (Komatsu *et al.*, 1993). It depends on the interference of light waves from the bottom surface of a crystal growing from a reflective substrate and the top surface, which is developing and, therefore, changes as a function of time with regard to its topological features. Because growth of a crystal surface is generally dominated by unique growth centres produced by dislocations or two-dimensional nuclei, the surfaces and the resultant interferograms change in a regular and periodic manner. Changes in the interferometric fringes with time provide accurate measures of the tangential and normal growth rates of a crystal (Vekilov *et al.*, 1993; Kuznetsov *et al.*, 1995; Kurihara *et al.*, 1996). From these data, one can determine the surface free energy and the kinetic coefficients that underlie the crystallization process.

EM (Durbin & Feher, 1990) and especially AFM (McPherson *et al.*, 2004) are powerful techniques for the investigation of crystallization mechanisms and their associated kinetics. The power of AFM lies in its ability to investigate crystal surfaces *in situ*, while they are still developing, thus permitting one to directly visualize, over time, the growth and change of a crystal face at near-nanometre resolution. The method is particularly useful in delineating the growth mechanisms involved, identifying dislocations, recording the kinetics of the changes and directly revealing impurity effects on the growth of protein crystals (Konnert *et al.*, 1994; Malkin *et al.*, 1996; Nakada *et al.*, 1999) (Fig. 4.1.5.1). AFM was also applied for the visualization of growth characteristics of crystals made of viruses (Malkin *et al.*, 1995) and RNA (Ng, Kuznetsov *et al.*, 1997). A noteworthy outcome of such studies was the sensitivity of growth to minor temperature changes. A variation of 2–3 °C was sufficient to transform the growth mechanism of yeast tRNA$^{\text{Phe}}$ from spiral screw dislocation growth at low supersaturation to two-dimensional island formation at high supersaturation (Ng, Kuznetsov *et al.*, 1997).

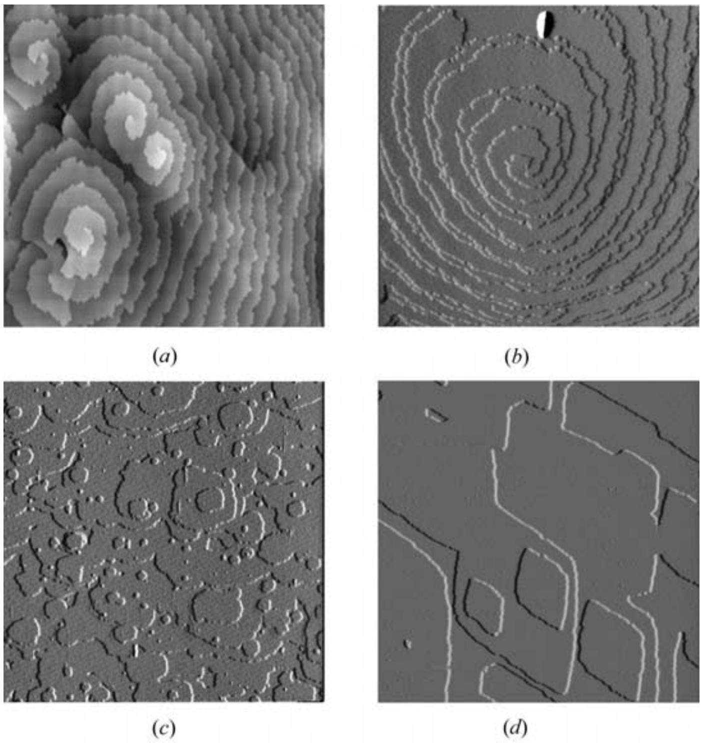


Figure 4.1.5.1
Growth mechanisms and visualization of protein crystal surfaces by AFM. In (*a*) and (*b*) are images of screw dislocations on the surfaces of crystals of the proteins canavalin and trypsin, respectively. The scan areas are 10 µm^2 in (*a*) and 30 µm^2 in (*b*). Screw dislocation growth predominates at low supersaturation. In (*c*) and (*d*) are examples of crystal growth by the formation of two-dimensional islands on the surfaces of crystals of the proteins thaumatin and glucose isomerase, respectively. The scan areas are 20 µm^2 in (*c*) and 11 µm^2 in (*d*). Growth by two-dimensional island formation and spread dominates at higher supersaturation.

4.1.5.3. Uncoupling nucleation and growth, and the constant-growth regime

The preparation of high-quality protein crystals should preferentially occur at lowest supersaturation and under a constant-growth regime. Achieving this aim with conventional crystallization methods is *a priori* not easy since growth of crystals is accompanied by a decrease of supersaturation in the crystallization medium (see the trajectory of crystal/solution equilibration in a phase diagram, Figs. 4.1.1.1 and 4.1.3.1). The implication is that it is possible to change the growth regime during the course of the crystallization process, as could be seen by AFM (Ng, Kuznetsov *et al.*, 1997). Such a change will perturb crystal formation and probably accounts for the frequently observed non-reproducibility in diffraction properties of protein crystals. Using flow cells with a constant supply of fresh protein may help to overcome this difficulty. Separating the nucleation and growth phases is another alternative (Chayen, 2005). This can be straightforwardly done by seeding procedures (see Section 4.1.4.4.).

4.1.5.4. Crystal perfection

The ultimate objective of structural biologists is to analyse crystals of high perfection, in other words, with a minimum of defects, disorder, impurity incorporation and internal stress. Such imperfections can be visualized using laser confocal microscopy combined with differential interference contrast microscopy (Iimura *et al.*, 2005; Sazaki *et al.*, 2005). They can also be evaluated by the resolution limit of diffraction, which expresses the average disorder of the molecules in the crystal lattice. Wilson plots provide good illustrations of diffraction quality for protein crystals. Other sources of disorder, such as the mosaic structure of the crystal, may strongly influence the quality of the diffraction data. They are responsible for increases in the diffuse background scatter and a broadening of diffraction intensities. These defects are difficult to monitor with precision, and dedicated techniques and instruments are required for accurate analysis.

Mosaicity can be defined experimentally by X-ray rocking width measurements. An overall diagnostic of crystal quality can be obtained by X-ray diffraction topography. Both techniques have been refined with lysozyme as a test case and were used for comparative analysis of crystals grown under different conditions, both on earth and in microgravity. For lysozyme and thaumatin, improvement of the mosaicity, as revealed by decreased rocking widths measured with synchrotron radiation, was observed for the microgravity-grown crystals (Snell *et al.*, 1995; Ng, Lorber *et al.*, 1997; Lorber, Sauter, Robert *et al.*, 1999).

An illustration of mosaic block character in a lysozyme crystal was provided by X-ray topography (Fourme *et al.*, 1995). Comparison of earth- and microgravity-grown lysozyme crystals showed a high density of defects in the earth control crystals, while in the microgravity case, several discrete regions were visible (Stojanoff *et al.*, 1996). X-ray topographs have also been used to compare crystal polymorphs (Izumi *et al.*, 1996), to monitor temperature-controlled growth of tetragonal lysozyme crystals (Stojanoff *et al.*, 1997), to study the effects of solution variations during growth on crystal perfection (Dobrianov *et al.*, 1998; Otálora *et al.*, 1999), to compare crystals grown in solution and in agarose gel (Lorber, Sauter, Ng *et al.*, 1999), and to map defects in the bulk of protein crystals (Hu *et al.*, 2001).

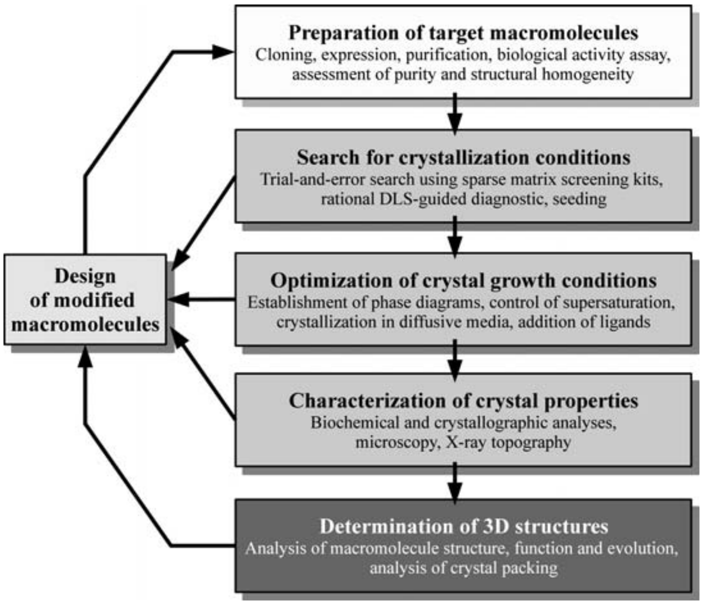

Preparation of target macromolecules
Cloning, expression, purification, biological activity assay, assessment of purity and structural homogeneity
Search for crystallization conditions
Trial-and-error search using sparse matrix screening kits, rational DLS-guided diagnostic, seeding
Design of modified macromolecules
Optimization of crystal growth conditions
Establishment of phase diagrams, control of supersaturation, crystallization in diffusive media, addition of ligands
Characterization of crystal properties
Biochemical and crystallographic analyses, microscopy, X-ray topography
Determination of 3D structures
Analysis of macromolecule structure, function and evolution, analysis of crystal packing

Figure 4.1.6.1

From the target molecule to its three-dimensional structure: a flowchart for a structure detemination. The figure illustrates the different steps the crystal grower will go through to access three-dimensional information. Ideally, the trajectory will be linear, crystallization remaining the major bottleneck in the whole process. If one gets stuck at any stage, the difficulties can often be circumvented by modifying the target molecule, either genetically or (bio)chemically, by improving its purity/homogeneity, or by adding ligands to stabilize its structure.

4.1.6. How to crystallize a new macromolecule: the structural biology view

Fig. 4.1.6.1 schematizes the progress of structural biology projects and highlights the pivotal role of the crystallization step. Despite tremendous progress in macromolecule and crystal preparation, crystallization remains the most serious bottleneck. Various strategies have been proposed to overcome the difficulty (*e.g.* D'Arcy, 1994; Ducruix & Giegé, 1999; Kundrot, 2004; McPherson, 2004; DeLucas *et al.*, 2005; Pusey *et al.*, 2005; Hughes & Ng, 2007; Li *et al.*, 2007; Chayen & Saridakis, 2008). The choice is largely dependent on the type of project [functional biology with well defined protein(s) or structural genomics with large ensembles of known and unknown proteins] and thus on non-exclusive rational, incomplete factorial screening, or empirical approaches.

4.1.6.1. How to start and how to choose what screening kits to start with

In practice, the very first questions posed when one begins to think about crystallization assays are: (i) which crystallization kit to start with, (ii) what to do if no crystals grow and (iii) what happens when one gets crystals. The answer to the first question is in fact simplest. One should start with a screen that does not unreasonably tax the supply of protein but which explores the widest volume of crystallization space, *i.e.* samples the largest number of crystallants and crystallant concentrations over the largest range of pH. There are many kits on the market using 96-sample trials that accomplish this well. If protein is severely limited, then there are 48-sample kits that do the job satisfactorily. In case of no success one should examine the resulting pattern of precipitates, phase separations, clear drops and assorted odd accumulations, and try to divine what might be the best option to try next. Good advice is to check how others have crystallized similar proteins and to try complexes of the protein with its physiological ligands and effectors.

If, on the other hand, one is fortunate and the first kit does indeed yield crystals, but perhaps of insufficient size or quality, or of troubling morphology, then crystallization must be optimized to get better ones. One should also check whether it is really the macromolecule that crystallized and not a ligand or any component of the crystallization mixture. Optimization means varying the chemical and physical parameters of the reagent mix that yielded the crystals and searching the crystallization parameter space by small increments away from the starting point. There is still a component of art and mystery in science, and this is one instance where its appreciation is paramount.

As a practical example, one can consider the 'Silver bullet' concept, which employs a limited set of fundamental crystallization conditions combined with a broad screen of potentially useful small-molecule additives (McPherson & Cudney, 2006), with the hypothesis that some of them might establish intermolecular noncovalent cross-links (Larson *et al.*, 2008). In the 'Uppsala' concept, a crystallization project entails two distinct steps: screening and optimization. The aim of the initial phase is to screen the parameters affecting crystallization as broadly as possible. If promising conditions are found, these are optimized with other protocols including grid and additive screens, seeding, and manipulation of drop kinetics (Bergfors, 2007).

4.1.6.2. Rules and general principles

Rules for sample preparation. The first concern is to obtain a macromolecular sample of the highest quality, second to collate all biochemical and biophysical information characterizing the macromolecule in order to design the best crystallization strategy, and finally, to establish precise protocols that ensure reproducibility of experiments. Solubility screening is an important issue and can be automated (Listwan *et al.*, 2009). Important as well is to clean and sterilize by filtration (over 0.22 µm porosity membranes, or even 0.10 µm) all solutions in contact with pure proteins, to remove dust and other solid particles, and to avoid contamination by microbes (Chayen, 2009). Inclusion of sodium azide in crystallizing solutions may discourage invasive bacteria and fungi. In vapour-diffusion assays, placing a small grain of thymol in the reservoir can prevent contamination. Thymol, however, can occasionally have specific effects on crystal growth (Chayen *et al.*, 1989) and thus may serve as an additive in screenings as well.

The preparation of solutions for crystallization experiments should follow some common rules. Stocks should be prepared with chemicals of the purest grade dissolved in double-distilled water, filtered through 0.22 µm membranes and ultracentrifuged (1 h at 100 000*g*) to remove aggregates. The chemical nature of the buffer is an important parameter, and the pH of buffers, which must be strictly controlled, is often temperature dependent, especially that of Tris buffers. Commercial PEG may contain ionic contaminants (Jurnak, 1986), or contaminants derived from peroxidation, and thus repurification is recommended (Ray & Puvathingal, 1985).

Mother liquors are defined as the solutions that contain all compounds (buffer, crystallant *etc.*) at the final concentration for crystallization, except the protein (to avoid dissolution of the crystals when transferred to a fresh mother liquor, soluble protein should be present). Samples of proteins often contain quantities of salt of unknown composition, and it is therefore wise to dialyse new batches against well characterized buffers. Whatever the crystallization method used, it almost always requires a high concentration of protein. This may imply concentration steps using devices operating under nitrogen pressure, by centrifugation or by lyophilization (note that lyophilization may denature proteins and that non-volatile salts will accumulate). Dialysis against high-molecular-weight PEG may also be used. During concentration, the pH and ionic strength may vary and, if not kept at the appropriate values, denaturation of samples may occur.

How to use phase diagrams and favour reproducibility. Crystallization requires bringing the protein to a supersaturated state that favours nucleation. Use of phase diagrams may be important for that purpose (Haas & Drenth, 1999; Sauter, Lorber *et al.*, 1999; Asherie, 2004; Lorber & Witz, 2008). If solubilities or phase diagrams are unavailable, it is nevertheless important to understand the correlation between solubility and the way supersaturation is reached in the different crystallization methods (Fig. 4.1.3.1). In dialysis, the protein concentration remains constant during equilibration. The initial concentration of the crystallant in the exterior solution leaves the macromolecule in an undersaturated state. With increasing concentration of this agent in the exterior solution, a state of supersaturation can be attained, leading to crystallization or precipitation. In a vapour-diffusion experiment, where the concentration of crystallant in the reservoir exceeds that in the drop, the protein will begin to concentrate from an undersaturated to a supersaturated state, with both protein and crystallant concentrations increasing. Crystals appear in the labile region. For crystals that appear first, the trajectory of equilibration is complex and the remaining concentration of protein in solution will converge towards a point located on the solubility curve. In batch crystallization using a closed vessel, three situations can occur: if the concentration of the protein is undersaturated, crystallization never occurs (unless another parameter such as temperature is varied); if it belongs to the supersaturated region between solubility and precipitation curves, crystals can grow until the remaining concentration of the protein in solution equals its solubility; if supersaturation is too high, the protein precipitates immediately, although in some cases crystals can grow from precipitates by Ostwald ripening (Ng *et al.*, 1996; Lorber & Witz, 2008). Altogether, the crystal habits (*e.g.* Sauter, Lorber *et al.*, 1999; Zhu *et al.*, 2001) and the underlying growth mechanisms (Fig. 4.1.5.1) are dependent on the region in the phase diagram where nucleation occurred and on the equilibration trajectories (Figs. 4.1.1.1 and 4.1.3.1).

Reproducibility should be a strategic concern (Newman *et al.*, 2007), but is often not reached despite the efforts of investigators to work with protein and crystallant solutions of identical chemical composition. Besides the many reasons that could explain this situation, the effect of the mixing protocol of the different components constituting a crystallization drop on the output of the experiment is often overlooked. Since mixing is not instantaneous, it will generate local supersaturation maxima and minima that can affect the crystallization process. Thus, when seeking reproducibility, active mixing of the crystallization drops is recommended, which is easy to perform when working manually but not trivial in automated procedures (Howard *et al.*, 2009).

4.1.6.3. Database mining and statistics

With the number of protein structures solved by X-ray crystallography now approaching 100 000, substantial databases of successful crystallization conditions and procedures have been developed (*e.g.* Peat *et al.*, 2005; Li *et al.*, 2007; Tung & Gallagher, 2009). From these it might be expected that some predictive insight would be obtained regarding the most probable crystallization conditions for proteins in general (*e.g.* Kimber *et al.*, 2003), and for specific families of proteins sharing common physical, chemical or functional properties such as aminoacyl-tRNA synthetases (Giegé *et al.*, 2008). However, the problem is that each protein remains an individual endowed with its own eccentricities, even within a family, and often these dramatically alter its crystallization behaviour. In addition, proteins in general may be exquisitely sensitive to only minor modifications to their properties, further complicating their rational classification in terms of crystal growth.

Nonetheless, statistical analyses are now being widely applied to the expanding databases and some results, both interesting and useful, are beginning to emerge. For example, reduced sets of the most favourable crystallization conditions have been proposed by several groups of investigators based on past successes. These may be useful because they reduce the number of crystallization trials in cases where either the amount of protein is limited or a very large number of constructs are to be screened. Favoured reagents and, particularly, useful additives have been identified by such analyses. Correlations have been sought between the physical or chemical properties of specific proteins and their manner of crystallization, such as between pI and crystallization pH, but this has had only limited success. A promising method generates relatively accurate predictions from protein composition and collocation of amino acids, pI and hydrophobicity deduced from sequence data, and number of solved structures of similar sequences (Mizianty & Kurgan, 2009). Interestingly, data mining of biophysical properties that control protein crystallization led to the conclusion that crystallization propensity depends primarily on the prevalence of well ordered surface epitopes capable of mediating interprotein interactions rather than on the overall thermodynamic stability of the proteins (Price *et al.*, 2009).

4.1.6.4. Strategic concerns: a summary

Homogeneity and stability. Perhaps the most important property of a system to be crystallized is its purity. Crystallization presupposes that identical units are available for incorporation into a periodic lattice. If crystallization fails, reconsidering purification protocols often helps achieve success. No homogeneous molecular population can remain so if its members alter their form, folding or association state. Hence, it is crucial that proteins in solution not be allowed to denature, aggregate or undergo conformational changes.

Solubility. Before a molecule can be crystallized, it must be solubilized. This means creation of monodisperse solutions free of aggregates and molecular clusters. Solubility and crystallizability strongly depend on substances (organic solvents and PEGs) that reduce the ionic strength of the solution (Papanikolau & Kokkinidis, 1997).

Supersaturation and nucleation. Crystals grow from systems displaced from equilibrium so that restoration requires formation of the solid state. Thus, the first task is to find ways to alter properties of the crystallizing solutions, such as by pH or temperature change, and to create supersaturated states. The number, size and quality of crystals depend on the mechanisms and rates of nuclei formation. For diffraction work in crystallization, one must seek to induce limited nucleation by adjustment of the physical and chemical properties of the system.

Association and variety. In forming crystals, molecules organize themselves through self-association to produce periodically repeating, three-dimensional arrays. Thus, it is necessary to facilitate positive molecular interactions while avoiding formation of precipitate, unspecific aggregates or phase separation. Proteins may crystallize under a wide spectrum of conditions and form many polymorphs. Thus, one should explore as many opportunities for crystallization as possible and explore the widest spectrum of biochemical, chemical and physical parameters.

Control and perfection. The ultimate value of any crystal is dependent on its perfection. Perturbations of the mother liquor are, in general, deleterious. Thus, crystallizing systems have to be maintained at an optimal state, without fluctuations or shock, until the crystals have matured. Crystallization conditions should be such as to favour crystal perfection, to minimize defects and high mosaicity of the growing crystals, and to minimize internal stress and incorporation of impurities. Predictions from crystal-growth theories may help to define such conditions (Chernov, 2003).

Impurities and preservation. Impurities can contribute to a failure to nucleate or to grow quality crystals. Thus, one must discourage their presence in the mother liquor and their incorporation into the lattice. Protein crystals may degrade and lose diffraction quality upon ageing. Thus, once grown, crystals may be stabilized by temperature change, addition of more crystallant or by some other suitable alteration in the mother liquor.

Rational design versus *automation*. When an investigation is focused on a single crystallographic objective, the structure solution of a specific protein or macromolecular complex, then it is wise to put as much time into headwork (thinking) as into handiwork (setting up crystallization samples). Examining and evaluating results, divining insights and setting new directions are certainly as important as actually dispensing samples into crystallization plates. There are, however, instances where it is more efficient to automate the process by which crystallization conditions are identified and optimized. This is true when there is not one, but many, real or potential targets.

4.1.7. The future of protein crystal growth

The fact that high-quality X-ray diffraction data can frequently be obtained from a single crystal of dimensions in the range of 20 to 50 μm has changed the objectives considerably from 40 years ago, when many crystals in the millimetre size range were required for a structure analysis. A consequence of this is that attention is turning increasingly from the systematic growth of large protein crystals (Bailey, 1942) to the nucleation and growth of any crystal. This direction has been further promoted by the development in the last 20 years of methods to use even the data from twinned or disordered crystals. One might be led to believe that the future of crystal growth is shrinking dramatically. Indeed, while smaller, fewer crystals are now the rule (with the exception of those required for neutron diffraction), this has not reduced the value of crystal perfection, nor the requirement that at least some sort of crystal be obtained. Thus, attention is now focused on nucleation, perhaps always the most problematic step in the crystallization process, and enhancing crystal perfection. These continue to remain formidable problems.

Furthermore, the objectives of crystallization, the entities to be crystallized, will continue to become more challenging. In addition to membrane proteins that pose difficult problems because of their solubility (see Chapter 4.2), interest has increasingly turned towards the solutions of the structures of RNA, glycoproteins (Chang *et al.*, 2007), lipoproteins, and larger protein or protein–nucleic acid complexes and assemblies. It is unlikely that crystals with unit cells much above 1200 Å can be solved with current X-ray technologies, but even those assemblies, such as large icosahedral viruses, that do yield crystals amenable to analysis are remarkably fragile in a mechanical sense, and the large unit-cell size requires that the crystals greatly exceed the small sizes of conventional protein crystals. Additional problems will arise from proteins conjugated with other entities of significant size such as lipids and oligosaccharides, which are often disordered, and with proteins that are unstructured, in whole or in part.

Finally, we have come to believe that the structure of a protein in the crystal is the same as the structure of the protein in solution. But when the protein has a spectrum of conformations in solution, as a consequence often of its function, then to visualize it in full one needs to see it in multiple crystal forms. Thus, it will be increasingly necessary to grow crystals not simply of the apo protein, but also of its possible ligand complexes, and possibly in several polymorphs. By studying the protein in a variety of crystal forms, its conformational variety may be appreciated and its dynamic range delineated.

We are indebted to Centre National de la Recherche Scientifique (CNRS) and Université de Strasbourg (UdS) for financial support.

References

Abe, F. & Horikoshi, K. (2001). *The biotechnological potential of piezophiles. Trends Biotechnol.* **19**, 102–108.

Adachi, H., Takano, K., Matsumura, H., Inoue, T., Mori, Y. & Sasaki, T. (2004). *Protein crystal growth with a two-liquid system and stirring solution. J. Synchrotron Rad.* **11**, 121–124.

Al-Haq, M. I., Lebrasseur, E., Choi, W. K., Tsuchiya, H., Torii, T., Yamazaki, H. & Shinohara, E. (2007). *An apparatus for electric-field-induced protein crystallization. J. Appl. Cryst.* **40**, 199–201.

Anand, K., Pal, D. & Hilgenfeld, R. (2002). *An overview on 2-methyl-2,4-pentanediol in crystallization and in crystals of biological macromolecules. Acta Cryst.* D**58**, 1722–1728.

Asherie, N. (2004). *Protein crystallization and phase diagrams. Methods*, **34,** 266–272.

Astier, J.-P. & Veesler, S. (2008). *Using temperature to crystallize proteins: a mini-review. Cryst. Growth Des.* **8**, 4200–4207.

Astier, J.-P., Veesler, S. & Boistelle, R. (1998). *Protein crystals orientation in a magnetic field. Acta Cryst.* D**54**, 703–706.

Ataka, M., Katoh, E. & Wakayama, N. I. (1997). *Magnetic orientation as a tool to study the initial stage of crystallization of lysozyme. J. Cryst. Growth*, **173**, 592–596.

Auerbach-Nevo, T., Zarivach, R., Peretz, M. & Yonath, A. (2005). *Reproducible growth of well diffracting ribosomal crystals. Acta Cryst.* D**61**, 713–719.

Bailey, K. (1942). *The growth of large single crystals of proteins. Faraday Soc. Trans.* **38**, 186–192.

Bancel, P. A., Cajipe, V. B., Rodier, F. & Witz, J. (1998). *Laser seeding for biomolecular crystallization. J. Cryst. Growth*, **191**, 537–544.

Bard, J., Ercolani, K., Svenson, K., Olland, A. & Somers, W. (2004). *Automated systems for protein crystallization. Methods*, **34**, 329–347.

Bergfors, T. (2003). *Seeds to crystals. J. Struct. Biol.* **142**, 66–76.

Bergfors, T. (2007). *Screening and optimization methods for nonautomated crystallization laboratories. Methods Mol. Biol.* **363**, 131–151.

Berry, I. M., Dym, O., Esnouf, R. M., Harlos, K., Meged, R., Perrakis, A., Sussman, J. L., Walter, T. S., Wilson, J. & Messerschmidt, A. (2006). *SPINE high-throughput crystallization, crystal imaging and recognition techniques: current state, performance analysis, new technologies and future aspects. Acta Cryst.* D**62**, 1137–1149.

Biertümpfel, C., Basquin, J., Birkenbihl, R. P., Suck, D. & Sauter, C. (2005). *Characterization of crystals of the Hjc resolvase from Archaeoglobus fulgidus grown in gel by counter-diffusion. Acta Cryst.* F**61**, 684–687.

Biertümpfel, C., Basquin, J., Suck, D. & Sauter, C. (2002). *Crystallization of biological macromolecules using agarose gel. Acta Cryst.* D**58**, 1657–1659.

Bolen, D. W. (2004). *Effects of naturally occurring osmolytes on protein stability and solubility: issues important in protein crystallization. Methods*, **34**, 312–322.

Bonneté, F., Malfois, M., Finet, S., Tardieu, A., Lafont, S. & Veesler, S. (1997). *Different tools to study interaction potentials in γ-crystallin solutions: relevance to crystal growth. Acta Cryst.* D**53**, 438–447.

Borgstahl, G. E. (2007). *How to use dynamic light scattering to improve the likelihood of growing macromolecular crystals. Methods Mol. Biol.* **363**, 109–129.

Boulot, G., Guillon, V., Mariuzza, R. A., Poljak, R. J., Riottot, M.-M., Souchon, H., Spinelli, S. & Tello, D. (1988). *Crystallization of antibody fragments and their complexes with antigen. J. Cryst. Growth*, **90**, 213–221.

Budayova-Spano, M., Dauvergne, F., Audiffren, M., Bactivelane, T. & Cusack, S. (2007). *A methodology and an instrument for the temperature-controlled optimization of crystal growth. Acta Cryst.* D**63**, 339–347.

Caffrey, M. & Cherezov, V. (2009). *Crystallizing membrane proteins using lipidic mesophases. Nat. Protoc.* **4**, 706–731.

Carson, M., Johnson, D. H., McDonald, H., Brouillette, C. & DeLucas, L. J. (2007). *His-tag impact on structure. Acta Cryst.* D**63**, 295–301.

Carter, C. W. Jr (1997). *Response surface methods for optimizing and improving reproducibility of crystal growth. Methods Enzymol.* **276**, 74–99.

Carter, C. W. & Carter, C. W. Jr (1979). *Protein crystallization using incomplete factorial experiments. J. Biol. Chem.* **254**, 12219–12223.

Carter, D. C., Lim, K., Ho, J. X., Wright, B. S., Twigg, P. D., Miller, T. Y., Chapman, J., Keeling, K., Ruble, J., Vekilov, P. G., Thomas, B. R., Rosenberger, F. & Chernov, A. A. (1999). *Lower dimer impurity incorporation may result in higher perfection of HEWL crystal grown in μg – a case study. J. Cryst. Growth*, **196**, 623–637.

Chang, V. E., Crispin, M., Aricescu, A. R., Harvey, D. J., Nettleship, J. E., Fennelly, J. A., Yu, C., Boles, K. S., Evans, E. J., Stuart, D. I., Dwek, R. A., Jones, E. Y., Owens, R. J. & Davis, S. J. (2007). *Glycoprotein structural genomics: solving the glycosylation problem. Structure*, **15**, 267–273.

Charron, C., Didierjean, C., Mangeot, J.-P. & Aubry, A. (2003). *The 'Octopus' plate for protein crystallization under an electric field. J. Appl. Cryst.* **36**, 1482–1483.

Charron, C., Robert, M.-C., Capelle, B., Kadri, A., Jenner, G., Giegé, R. & Lorber, B. (2002). *X-ray diffraction properties of protein crystals prepared in agarose gel under hydrostatic pressure. J. Cryst. Growth*, **245**, 321–333.

Chayen, N. E. (1996). *A novel technique for containerless protein crystallization. Protein Eng.* **9**, 927–929.

Chayen, N. E. (1997). *A novel technique to control the rate of vapor diffusion, giving larger protein crystals. J. Appl. Cryst.* **30**, 198–202.

Chayen, N. E. (2005). *Methods for separating nucleation and growth in protein crystallisation. Prog. Biophys. Mol. Biol.* **88**, 329–337.

Chayen, N. E. (2009). *Rigorous filtration for protein crystallization. J. Appl. Cryst.* **42**, 743–744.

Chayen, N. E., Lloyd, L. F., Collyer, C. A. & Blow, D. M. (1989). *Trigonal crystals of glucose isomerase require thymol for their growth and stability. J. Cryst. Growth*, **97**, 367–374.

Chayen, N. E. & Saridakis, E. (2008). *Protein crystallization: from purified protein to diffraction-quality crystal. Nat. Methods*, **5**, 147–153.

Chayen, N. E., Saridakis, E. & Sear, R. P. (2006). *Experiment and theory for heterogeneous nucleation of protein crystals in a porous medium. Proc. Natl Acad. Sci. USA*, **103**, 597–601.

Chayen, N. E., Shaw Stewart, P. D., Maeder, D. L. & Blow, D. M. (1990). *An automated system for microbatch protein crystallization and screening. J. Appl. Cryst.* **23**, 297–302.

Chen, X. W., Ji, Y. P. & Wang, J. H. (2010). *Improvement on the crystallization of lysozyme in the presence of hydrophilic ionic liquid. Analyst*, **135**, 2241–2248.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K. & Stevens, R. C. (2007). *High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science*, **318**, 1258–1265.

Chernov, A. A. (2003). *Protein crystals and their growth. J. Struct. Biol.* **142**, 3–21.

Christopher, G. K., Phipps, A. G. & Gray, R. J. (1998). *Temperature-dependent solubility of selected proteins. J. Cryst. Growth*, **191**, 820–826.

Collins, K. D. (2004). *Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process. Methods*, **34**, 300–311.

Crosio, M.-P. & Jullien, M. (1992). *Fluorescence study of precrystallization of ribonuclease A: effects of salts. J. Cryst. Growth*, **122**, 66–70.

Cudney, B., Patel, S. & McPherson, A. (1994). *Crystallization of macromolecules in silica gels. Acta Cryst.* D**50**, 479–483.

Dale, G. E., Oefner, C. & D'Arcy, A. (2003). *The protein as a variable in protein crystallization. J. Struct. Biol.* **142**, 88–97.

D'Arcy, A. (1994). *Crystallizing proteins – a rational approach? Acta Cryst.* D**50**, 469–471.

D'Arcy, A., Elmore, C., Stihle, M. & Johnston, J. E. (1996). *A novel approach to crystallizing proteins under oil. J. Cryst. Growth*, **168**, 175–180.

D'Arcy, A., MacSweeney, A. & Haber, A. (2003). *Using natural seeding material to generate nucleation in protein crystallization experiments. Acta Cryst.* D**59**, 1343–1346.

D'Arcy, A., MacSweeney, A., Stihle, M. & Haber, A. (2003). *The advantages of using a modified microbatch method for rapid screening of protein crystallization conditions. Acta Cryst.* D**59**, 396–399.

D'Arcy, A., Sweeney, A. M. & Haber, A. (2004). *Practical aspects of using the microbatch method in screening conditions for protein crystallization. Methods*, **34**, 323–328.

D'Arcy, A., Villard, F. & Marsh, M. (2007). *An automated microseed matrix-screening method for protein crystallization. Acta Cryst.* D**63**, 550–554.

Declercq, J.-P., Evrard, C., Carter, D. C., Wright, B. S., Etienne, G. & Parello, J. (1999). *A crystal of a typical EF-hand protein grown under microgravity diffracts X-rays beyond 0.9 Å resolution. J. Cryst. Growth*, **196**, 595–601.

DeLucas, L. J., Bray, T. L., Nagy, L., McCombs, D., Chernov, N., Hamrick, D., Cosenza, L., Belgovskiy, A., Stoops, B. & Chait, A. (2003). *Efficient protein crystallization. J. Struct. Biol.* **142**, 188–206.

DeLucas, L. J., Hamrick, D., Cosenza, L., Nagy, L., McCombs, D., Bray, T., Chait, A., Stoops, B., Belgovskiy, A., Wilson, W. W., Parham, M. & Chernov, N. (2005). *Protein crystallization: virtual screening and optimization. Prog. Biophys. Mol. Biol.* **88**, 285–309.

DeLucas, L. J., Long, M. M., Moore, K. M., Rosenblum, W. M., Bray, T. L., Smith, C., Carson, M., Narayana, S. V. L., Harrington, M. D., Carter, D., Clark, A. D., Nanni, R. G., Ding, J., Jacobomolina, A., Kamer, G., Hughes, S. H., Arnold, E., Einspahr, H. M., Clancy, L. L., Rao, G. S. J., Cook, P. F., Harris, B. G., Munson, S. H., Finzel, B. C., McPherson, A., Weber, P. C., Lewandowski, F. A., Nagabhushan, T. L., Trotta, P. P., Reichert, P., Navia, M. A., Wilson, K. P., Thomson, J. A., Richards, R. N., Bowersox, K. D., Meade, C. J., Baker, E. S., Bishop, S. P., Dunbar, B. J., Trinh, E., Prahl, J., Sacco, A. & Bugg, C. E. (1994). *Recent results and new hardware developments for protein crystal growth in microgravity. J. Cryst. Growth*, **135**, 183–195.

DeMattei, R. C. & Feigelson, R. S. (1993). *Thermal methods for crystallizing biological macromolecules. J. Cryst. Growth*, **128**, 1225–1231.

Derewenda, Z. S. (2004). *The use of recombinant methods and molecular engineering in protein crystallization. Methods*, **34**, 354–363.

Dhouib, K., Khan Malek, C., Pfleging, W., Gauthier-Manuel, B., Duffait, R., Thuillier, G., Ferrigno, R., Jacquamet, L., Ohana, J., Ferrer, J.-L., Théobald-Dietrich, A., Giegé, R., Lorber, B. & Sauter, C. (2009). *Microfluidic chips for the crystallization of biomacromolecules by counter-diffusion and on-chip crystal X-ray analysis. Lab Chip*, **9**, 1412–1421.

Dierks, K., Meyer, A., Einspahr, H. & Betzel, C. (2008). *Dynamic light scattering in protein crystallization droplets: adaptations for analysis and optimization of crystallization processes. Cryst. Growth Des.* **8**, 1628–1634.

Dobrianov, I., Finkelstein, K. D., Lemay, S. G. & Thorne, R. E. (1998). *X-ray topographic studies of protein crystal perfection and growth. Acta Cryst.* D**54**, 922–937.

Dock, A.-C., Lorber, B., Moras, D., Pixa, G., Thierry, J.-C. & Giegé, R. (1984). *Crystallization of transfer ribonucleic acids. Biochimie*, **66**, 179–201.

Dock-Bregeon, A.-C., Chevrier, B., Podjarny, A., Moras, D., deBear, J. S., Gough, G. R., Gilham, P. T. & Johnson, J. E. (1988). *High resolution structure of the RNA duplex* $[U(U\text{-}A)6A]_2$*. Nature (London)*, **335**, 375–378.

Ducruix, A. & Giegé, R. (1999). Editors. *Crystallization of Nucleic Acids and Proteins: a Practical Approach*, 2nd ed. Oxford: IRL Press.

Durbin, S. D. & Feher, G. (1990). *Studies of crystal growth mechanisms by electron microscopy. J. Mol. Biol.* **212**, 763–774.

Durbin, S. D. & Feher, G. (1996). *Protein crystallization. Annu. Rev. Phys. Chem.* **47**, 171–204.

Ebel, C., Faou, P. & Zaccaï, G. (1999). *Protein–solvent and weak protein–protein interactions in halophilic malate dehydrogenase. J. Cryst. Growth*, **196**, 395–402.

Ferré-D'Amaré, A. R. (2010). *Use of the spliceosomal protein U1A to facilitate crystallization and structure determination of complex RNAs. Methods*, **52**, 159–167.

Ferré-D'Amaré, A. R. & Burley, S. (1997). *Dynamic light scattering in evaluating crystallizability of macromolecules. Methods Enzymol.* **274**, 157–166.

Finet, S., Bonneté, F., Frouin, J., Provost, K. & Tardieu, A. (1998). *Lysozyme crystal growth, as observed by small angle X-ray scattering, proceeds without crystallization intermediates. Eur. J. Biophys.* **27**, 263–271.

Forsythe, E., Achari, A. & Pusey, M. L. (2006). *Trace fluorescent labeling for high-throughput crystallography. Acta Cryst.* D**62**, 339–346.

Fourme, R., Ducruix, A., Riès-Kautt, M. & Capelle, B. (1995). *The perfection of protein crystals probed by direct recording of Bragg reflection profiles with a quasi-planar X-ray wave. J. Synchrotron Rad.* **2**, 136–142.

Frontana-Uribe, B. A. & Moreno, A. (2008). *On electrochemically assisted protein crystallization and related methods. Cryst. Growth Des.* **8**, 4194–4199.

Garcia-Ruiz, J. M. (2003). *Counterdiffusion methods for macromolecular crystallization. Methods Enzymol.* **368**, 130–154.

Garcia-Ruiz, J. M. & Moreno, A. (1994). *Investigations on protein crystal growth by the gel acupuncture method. Acta Cryst.* D**50**, 484–490.

Garcia-Ruiz, J. M., Moreno, A., Otalora, F., Rondon, D., Viedma, C. & Zauscher, F. (1998). *Teaching protein crystallization by the gel acupuncture method. J. Chem. Educ.* **75**, 442–446.

Garcia-Ruiz, J. M., Novella, M. L. & Otalora, F. (1999). *Supersaturation patterns in counter-diffusion crystallization methods followed by Mach–Zehnder interferometry. J. Cryst. Growth*, **196**, 703–710.

Gavira, J. A. & Garcia-Ruiz, J. M. (2002). *Agarose as crystallisation media for proteins II: trapping of gel fibres into the crystals. Acta Cryst.* D**58**, 1653–1656.

Gavira, J. A., Toh, D., Lopez-Jaramillo, J., Garcia-Ruiz, J. M. & Ng, J. D. (2002). *Ab initio crystallographic structure determination of insulin from protein to electron density without crystal handling. Acta Cryst.* D**58**, 1147–1154.

Georgalis, Y., Zouni, A., Eberstein, W. & Saenger, W. (1993). *Formation dynamics of protein precrystallization fractal clusters. J. Cryst. Growth*, **126**, 245–260.

George, A., Chiang, Y., Guo, B., Abrabshari, A., Cai, Z. & Wilson, W. W. (1997). *Second virial coefficient as predictor in protein crystal growth. Methods Enzymol.* **276**, 100–110.

Giegé, R., Dock, A.-C., Kern, D., Lorber, B., Thierry, J.-C. & Moras, D. (1986). *The role of purification in the crystallization of proteins and nucleic acids. J. Cryst. Growth*, **76**, 554–561.

Giegé, R., Drenth, J., Ducruix, A., McPherson, A. & Saenger, W. (1995). *Crystallogenesis of biological macromolecules. Biological, microgravity, and other physico-chemical aspects. Prog. Cryst. Growth Charact.* **30**, 237–281.

Giegé, R., Moras, D. & Thierry, J.-C. (1977). *Yeast transfer* RNA^{Asp}*: a new high resolution X-ray diffracting crystal form of a transfer RNA. J. Mol. Biol.* **115**, 91–96.

Giegé, R., Touzé, E., Lorber, B., Théobald-Dietrich, A. & Sauter, C. (2008). *Crystallogenesis trends of free and liganded aminoacyl-tRNA synthetases. Cryst. Growth Des.* **8**, 4297–4306.

Girard, E., Kahn, R., Mezouar, M., Dhaussy, A. C., Lin, T., Johnson, J. E. & Fourme, R. (2005). *The first crystal structure of a macromolecular assembly under high pressure: CpMV at 330 MPa. Biophys. J.* **88**, 3562–3571.

Goldberg, M. E., Expert-Bezancon, N., Vuillard, L. & Rabilloud, T. (1996). *Non-detergent sulphobetaines: a new class of molecules that facilitate in vitro protein renaturation. Fold Des.* **1**, 21–27.

Golden, B. L. & Kundrot, C. E. (2003). *RNA crystallization. J. Struct. Biol.* **142**, 98–107.

Gonzalez-Ramirez, L. A., Caballero, A. G. & Garcia-Ruiz, J. M. (2008). *Investigation of the compatibility of gels with precipitating agents and detergents in protein crystallization experiments. Cryst. Growth Des.* **8**, 4291–4296.

Gonzalez-Ramirez, L. A., Carrera, J., Gavira, J. A., Melero-Garcia, E. & Garcia-Ruiz, J. M. (2008). *Granada Crystallization Facility-2: a versatile platform for crystallization in space. Cryst. Growth Des.* **8**, 4324–4329.

Gosavi, R. A., Bhamidi, V., Varanasio, S. & Schall, C. A. (2009). *Beneficial effect of solubility enhancers on protein crystal nucleation and growth. Langmuir*, **25**, 4579–4587.

Green, A. A. & Hughes, W. L. (1955). *Protein fractionation on the basis of the solubility in aqueous solutions of salts and organic solvents. Methods Enzymol.* **1**, 67–90.

Gripon, C., Legrand, L., Rosenman, I., Vidal, O., Robert, M.-C. & Boué, F. (1997). *Lysozyme–lysozyme interactions in under- and super-saturated solutions: a simple relation between the second virial coefficients in* H_2O *and* D_2O*. J. Cryst. Growth*, **178**, 575–584.

Guo, M., Shapiro, R., Schimmel, P. & Yang, X.-L. (2010). *Introduction of a leucine half-zipper engenders multiple high-quality crystals of a recalcitrant tRNA synthetase. Acta Cryst.* D**66**, 243–250.

Haas, C. & Drenth, J. (1999). *Understanding protein crystallization on the basis of the phase diagram. J. Cryst. Growth*, **196**, 388–394.

Hammadi, Z., Astier, J. P., Morin, R. & Veesler, S. (2009). *Spatial and temporal control of nucleation by localized DC electric field. Cryst. Growth Des.* **9**, 3346–3347.

Hampel, M., Labananskas, P. G., Conners, L., Kirkegard, U. L., RajBhandary, U. L., Sigler, P. B. & Bock, R. M. (1968). *Single crystals of transfer RNA from formylmethionine and phenylalanine transfer RNAs*. *Science*, **162**, 1384–1386.

Hansen, C. L., Classen, S., Berger, J. M. & Quake, S. R. (2006). *A microfluidic device for kinetic optimization of protein crystallization and in situ structure determination*. *J. Am. Chem. Soc.* **128**, 3142–3143.

Hansen, C. L., Skordalakes, E., Berger, J. M. & Quake, S. R. (2002). *A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion*. *Proc. Natl Acad. Sci. USA*, **99**, 16531–16536.

Hansen, C. L., Sommer, M. O. & Quake, S. R. (2004). *Systematic investigation of protein phase behavior with a microfluidic formulator*. *Proc. Natl Acad. Sci. USA*, **101**, 14431–14436.

Harlos, K. (1992). *Micro-bridges for sitting-drop crystallizations*. *J. Appl. Cryst.* **25**, 536–538.

Helliwell, J. R. & Chayen, N. E. (2007). *Crystallography: a down-to-Earth approach*. *Nature (London)*, **448**, 658–659.

Henisch, H. K. (1988). *Crystals in Gels and Liesegang Rings*. Cambridge, MA: Cambridge University Press.

Hirschler, J., Charon, M.-H. & Fontecilla-Camps, J. C. (1995). *The effects of filtration on protein nucleation in different growth media*. *Protein Sci.* **4**, 2573–2577.

Holbrook, S. R., Holbrook, E. L. & Walukiewicz, H. E. (2001). *Crystallization of RNA*. *CMLS Cell. Mol. Life Sci.* **58**, 234–243.

Hosfield, D., Palan, J., Hilgers, M., Scheibe, D., McRee, D. E. & Stevens, R. C. (2003). *A fully integrated protein crystallization platform for small-molecule drug discovery*. *J. Struct. Biol.* **142**, 207–217.

Howard, E. I., Fernandez, J. M. & Garcia-Ruiz, J. M. (2009). *On the mixing of protein crystallization cocktails*. *Cryst. Growth Des.* **9**, 2707–2712.

Hu, Z. W., Thomas, B. R. & Chernov, A. A. (2001). *Laboratory multiple-crystal X-ray topography and reciprocal-space mapping of protein crystals: influence of impurities on crystal perfection*. *Acta Cryst.* D**57**, 840–846.

Hughes, R. C. & Ng, J. D. (2007). *Can small laboratories do structural genomics?* *Cryst. Growth Des.* **7**, 2226–2238.

Iimura, Y., Yoshizaki, I., Yoda, S. & Komatsu, H. (2005). *Development of an embedding method for analyzing the impurity distribution in protein crystals*. *Cryst. Growth Des.* **5**, 295–300.

Izumi, K., Sawamura, S. & Ataka, M. (1996). *X-ray topography of lysozyme crystals*. *J. Cryst. Growth*, **168**, 106–111.

Jakoby, W. B. (1971). *Crystallization as a purification technique*. *Methods Enzymol.* **22**, 248–252.

Jancarik, J. & Kim, S.-H. (1991). *Sparse matrix sampling: a screening method for crystallization of proteins*. *J. Appl. Cryst.* **24**, 409–411.

Jeruzalmi, D. & Steitz, T. A. (1997). *Use of organic cosmotropic solutes to crystallize flexible proteins: application to T7 RNA polymerase and its complex with the inhibitor T7 lysozyme*. *J. Mol. Biol.* **274**, 748–756.

Judge, R. A., Forsythe, E. L. & Pusey, M. L. (1998). *The effect of protein impurities on lysozyme crystal growth*. *Biotechnol. Bioeng.* **59**, 777–785.

Judge, R. A., Jacobs, R. S., Frazier, T., Snell, E. H. & Pusey, M. L. (1999). *The effect of temperature and solution pH on the nucleation of tetragonal lysozyme crystals*. *Biophys. J.* **77**, 1585–1593.

Judge, R. A., Takahashi, S., Longenecker, K. L., Fry, E. H., Abad-Zapatero, C. & Chiu, M. L. (2009). *The effect of ionic liquids on protein crystallization and X-ray diffraction studies*. *Cryst. Growth Des.* **9**, 3463–3469.

Jurnak, F. (1986). *Effect of chemical impurities in polyethylene glycol on macromolecular crystallization*. *J. Cryst. Growth*, **76**, 577–582.

Kadri, A., Jenner, G., Damak, M., Lorber, B. & Giegé, R. (2003). *Crystallogenesis studies of proteins in agarose gel – combined effect of high hydrostatic pressure and pH*. *J. Cryst. Growth*, **257**, 390–402.

Kakinouchi, K., Adachi, H., Matsumura, H., Inoue, T., Murakami, S., Mori, Y., Koga, Y., Takano, K. & Kanaya, S. (2006). *Effect of ultrasonic irradiation on protein crystallization*. *J. Cryst. Growth*, **292**, 437–440.

Kam, Z., Shore, H. B. & Feher, G. (1978). *On the crystallization of proteins*. *J. Mol. Biol.* **123**, 539–555.

Karpukhina, S. Y., Barynin, V. V. & Lobanova, G. M. (1975). *Crystallization of catalase in the ultracentrifuge*. *Sov. Phys. Crystallogr.* **20**, 417–418.

Khurshid, S., Haire, L. F. & Chayen, N. E. (2010). *Automated seeding for the optimization of crystal quality*. *J. Appl. Cryst.* **43**, 752–756.

Kim, Y., Quartey, P., Li, H., Volkart, L., Hatzos, C., Chang, C., Nocek, B., Cuff, M., Osipiuk, J., Tan, K., Fan, Y., Bigelow, L., Maltseva, N., Wu, R., Borovilos, M., Duggan, E., Zhou, M., Binkowski, T. A., Zhang, R. G. & Joachimiak, A. (2008). *Large-scale evaluation of protein reductive methylation for improving protein crystallization*. *Nat. Methods*, **5**, 853–854.

Kimber, M. S., Vallee, F., Houston, S., Necakov, A., Skarina, T., Evdokimova, E., Beasley, S., Christendat, D., Savchenko, A., Arrowsmith, C. H., Vedada, M., Gerstein, M. & Edwards, A. M. (2003). *Data mining crystallization databases: knowledge-based approaches to optimize protein crystal screens*. *Proteins*, **51**, 562–568.

Kimble, W. L., Paxton, T. E., Rousseau, R. W. & Sambanis, A. (1998). *The effect of mineral substrates on the crystallization of lysozyme*. *J. Cryst. Growth*, **187**, 268–276.

Koide, S. (2009). *Engineering of recombinant crystallization chaperones*. *Curr. Opin. Struct. Biol.* **19**, 449–457.

Komatsu, H., Miyashita, S. & Suzuki, Y. (1993). *Interferometric observation of the interfacial concentration gradient layers around a lysozyme crystal*. *Jpn. J. Appl. Phys.* **32**, L1855–L1857.

Konnert, J. H., D'Antonio, P. & Ward, K. B. (1994). *Observation of growth steps, spiral dislocations and molecular packing on the surface of lysozyme crystals with the atomic force microscope*. *Acta Cryst.* D**50**, 603–613.

Koszelak, S. & McPherson, A. (1988). *Time lapse microphotography of protein crystal growth using a color VCR*. *J. Cryst. Growth*, **90**, 340–343.

Koszelak, S., Martin, D., Ng, J. & McPherson, A. (1991). *Protein crystal growth rates determined by time-lapse microphotography*. *J. Cryst. Growth*, **110**, 177–181.

Kundrot, C. E. (2004). *Which strategy for a protein crystallization project?* *CMLS Cell. Mol. Life Sci.* **61**, 525–536.

Kundrot, C. E., Judge, R. A., Pusey, M. L. & Snell, E. H. (2001). *Microgravity and macromolecular crystallography*. *Cryst. Growth Des.* **1**, 87–99.

Kurihara, K., Miyashita, S., Sazaki, G., Nakada, T., Suzuki, Y. & Komatsu, H. (1996). *Interferometric study on the crystal growth of tetragonal lysozyme crystals*. *J. Cryst. Growth*, **166**, 904–906.

Kutá Smatanová, I., Gavira, J. A., Rezacova, P., Vacha, F. & Garcia-Ruiz, J. M. (2006). *New techniques for membrane protein crystallization tested on photosystem II core complex of Pisum sativum*. *Photosynth. Res.* **90**, 255–259.

Kuznetsov, Y. G., Day, J., Newman, R. & McPherson, A. (2000). *Chimeric human-simian anti-CD4 antibodies form crystalline high symmetry particles*. *J. Struct. Biol.* **131**, 108–115.

Kuznetsov, Y. G., Malkin, A. J., Greenwood, A. & McPherson, A. (1995). *Interferometric studies of growth kinetics and surface morphology in macromolecular crystal growth: canavalin, thaumatin, and turnip yellow mosaic virus*. *J. Struct. Biol.* **114**, 184–196.

Kuznetsov, Y. G., Malkin, A. J. & McPherson, A. (2001). *The influence of precipitant concentration on macromolecular crystal growth mechanisms*. *J. Cryst. Growth*, **232**, 114–118.

Lange, C., Patil, G. & Rudolph, R. (2005). *Ionic liquids as refolding additives: N′-alkyl and N′-(omega-hydroxyalkyl) N-methylimidazolium chlorides*. *Protein Sci.* **14**, 2693–2701.

Larson, S. B., Day, J. S., Cudney, R. & McPherson, A. (2007). *A novel strategy for the crystallization of proteins: X-ray diffraction validation*. *Acta Cryst.* D**63**, 310–318.

Larson, S. B., Day, J. S., Nguyen, C., Cudney, R. & McPherson, A. (2008). *Progress in the development of an alternative approach to macromolecular crystallization*. *Cryst. Growth Des.* **8**, 3038–3052.

Lee, I. S., Evans, J. M. B., Erdemir, D., Lee, A. Y., Garetz, B. A. & Myerson, A. S. (2008). *Nonphotochemical laser induced nucleation of hen egg white lysozyme crystals*. *Cryst. Growth Des.* **8**, 4255–4261.

Lee, S. S. J. & Cudney, R. (2004). *A modified microdialysis button for use in protein crystallization*. *J. Appl. Cryst.* **37**, 504–505.

Lenhoff, A. M., Pjura, P. E., Dilmore, J. G. & Godlewski, T. S. Jr (1997). *Ultracentrifugal crystallization of proteins: transport-kinetic modelling and experimental behavior of catalase*. *J. Cryst. Growth*, **180**, 113–126.

Li, C. M., Kirkwood, K. L. & Brayer, G. D. (2007). *The biological crystallization resource: facilitating knowledge-based protein crystallizations*. *Cryst. Growth Des.* **7**, 2147–2152.

Lin, T., Schildkamp, W., Brister, K., Doerschuk, P. C., Somayazulu, M., Mao, H. K. & Johnson, J. E. (2005). *The mechanism of high-pressure-induced ordering in a macromolecular crystal*. *Acta Cryst.* D**61**, 737–743.

Lin, Y. B., Zhu, D. W., Wang, T., Song, J., Zou, Y. S., Zhang, Y. L. & Lin, S. X. (2008). *An extensive study of protein phase diagram modification:*

increasing macromolecular crystallizability by temperature screening. *Cryst. Growth Des.* **8**, 4277–4283.

Listwan, P., Terwilliger, T. C. & Waldo, G. S. (2009). *Automated, high-throughput platform for protein solubility screening using a split-GFP system*. *J. Struct. Funct. Genom.* **10**, 47–55.

Lorber, B. (2008). *Virus and protein crystallization under hypergravity*. *Cryst. Growth Des.* **8**, 2964–2969.

Lorber, B. & Giegé, R. (1992). *A versatile reactor for temperature controlled crystallization of biological macromolecules*. *J. Cryst. Growth*, **122**, 168–175.

Lorber, B. & Giegé, R. (1996). *Containerless protein crystallization in floating drops: application to crystal growth monitoring under reduced nucleation conditions*. *J. Cryst. Growth*, **168**, 204–215.

Lorber, B., Jenner, G. & Giegé, R. (1996). *Effect of high hydrostatic pressure on the nucleation and growth of protein crystals*. *J. Cryst. Growth*, **158**, 103–117.

Lorber, B., Ng, J. D., Lautenschlager, P. & Giegé, R. (2000). *Growth kinetics and motion of thaumatin crystals during USML-2 and LMS microgravity missions and comparison with earth controls*. *J. Cryst. Growth*, **208**, 665–677.

Lorber, B., Sauter, C., Ng, J. D., Zhu, D.-W., Giegé, R., Vidal, O., Robert, M.-C. & Capelle, B. (1999). *Characterization of protein and virus crystals by quasi-planar wave X-ray topography: a comparison between crystals grown in solution and in agarose gel*. *J. Cryst. Growth*, **204**, 357–368.

Lorber, B., Sauter, C., Robert, M.-C., Capelle, B. & Giegé, R. (1999). *Crystallization within agarose gel in microgravity improves the quality of thaumatin crystals*. *Acta Cryst.* D**55**, 1491–1494.

Lorber, B., Sauter, C., Théobald-Dietrich, A., Moreno, A., Schellenberger, P., Robert, M.-C., Capelle, B., Sanglier, S., Potier, N. & Giegé, R. (2009). *Crystal growth of proteins, nucleic acids, and viruses in gels*. *Prog. Biophys. Mol. Biol.* **101**, 13–25.

Lorber, B. & Witz, J. (2008). *An investigation of the crystallogenesis of an icosahedral RNA plant virus with solubility phase diagrams*. *Cryst. Growth Des.* **8**, 1522–1529.

Lübbert, D., Meents, A. & Weckert, E. (2004). *Accurate rocking-curve measurements on protein crystals grown in a homogeneous magnetic field of 2.4 T*. *Acta Cryst.* D**60**, 987–998.

Luft, J. & Cody, V. (1989). *A simple capillary vapor diffusion apparatus for surveying macromolecular crystallization conditions*. *J. Appl. Cryst.* **22**, 396.

Luft, J. & De Titta, G. T. (1992). *HANGMAN: a macromolecular hanging-drop vapor-diffusion technique*. *J. Appl. Cryst.* **25**, 324–325.

Luft, J. R., Albright, D. T., Baird, J. K. & DeTitta, G. T. (1996). *The rate of water equilibration in vapor-diffusion crystallization. Dependence on the distance from the droplet to the reservoir*. *Acta Cryst.* D**52**, 1098–1106.

Luft, J. R., Collins, R. J., Fehrman, N. A., Lauricella, A. M., Veatch, C. K. & DeTitta, G. T. (2003). *A deliberate approach to screening for initial crystallization conditions of biological macromolecules*. *J. Struct. Biol.* **142**, 170–179.

Luft, J. R., Rak, D. M. & DeTitta, G. T. (1999*a*). *Microbatch macromolecular crystallization in micropipettes*. *J. Cryst. Growth*, **196**, 450–455.

Luft, J. R., Rak, D. M. & DeTitta, G. T. (1999*b*). *Microbatch macromolecular crystallization on a thermal gradient*. *J. Cryst. Growth*, **196**, 447–449.

Maclean, D. S., Qian, Q. & Middaugh, C. R. (2002). *Stabilization of proteins by low molecular weight multi-ions*. *J. Pharm. Sci.* **91**, 2220–2229.

McPherson, A. (1976). *Crystallization of proteins from polyethylene glycol*. *J. Biol. Chem.* **251**, 6300–6303.

McPherson, A. (1982). *The Preparation and Analysis of Protein Crystals*. New York: Wiley.

McPherson, A. (1990). *Current approaches to macromolecular crystallization*. *Eur. J. Biochem.* **189**, 1–23.

McPherson, A. (1991). *A brief history of protein crystal growth*. *J. Cryst. Growth*, **110**, 1–10.

McPherson, A. (1996). *Macromolecular crystal growth in microgravity*. *Crystallogr. Rev.* **6**, 157–305.

McPherson, A. (1999). *Crystallization of Biological Macromolecules*. Cold Spring Harbor and New York: Cold Spring Harbor Laboratory Press.

McPherson, A. (2001). *A comparison of salts for the crystallization of macromolecules*. *Protein Sci.* **10**, 418–422.

McPherson, A. (2004). *Introduction to protein crystallization*. *Methods*, **34**, 254–265.

McPherson, A. & Cudney, B. (2006). *Searching for silver bullets: an alternative strategy for crystallizing macromolecules*. *J. Struct. Biol.* **156**, 387–406.

McPherson, A. & Giegé, R. (2007). *Crystallogenesis research for biology in the last two decades as seen from the international conferences on the crystallization of biological macromolecules*. *Cryst. Growth Des.* **7**, 2126–2133.

McPherson, A., Kuznetsov, Y. G., Malkin, A. J. & Plomp, M. (2004). *Macromolecular crystal growth investigations using atomic force microscopy*. *J. Synchrotron Rad.* **11**, 21–23.

McPherson, A., Malkin, A. J. & Kuznetsov, Y. G. (1995). *The science of macromolecular crystallization*. *Structure*, **3**, 759–768.

McPherson, A., Malkin, A. J., Kuznetsov, Y. G. & Koszelak, S. (1996). *Incorporation of impurities into macromolecular crystals*. *J. Cryst. Growth*, **168**, 74–92.

McPherson, A. & Shlichta, P. (1988*a*). *Heterogeneous and epitaxial nucleation of protein crystals on mineral surfaces*. *Science*, **239**, 385–387.

McPherson, A. & Shlichta, P. (1988*b*). *The use of heterogeneous and epitaxial nucleants to promote the growth of protein crystals*. *J. Cryst. Growth*, **90**, 47–50.

Malkin, A. J., Kuznetsov, Y. G., Land, T. A., DeYoreo, J. J. & McPherson, A. (1995). *Mechanisms of growth for protein and virus crystals*. *Nat. Struct. Biol.* **2**, 956–959.

Malkin, A. J., Kuznetsov, Y. G. & McPherson, A. (1996). *Defect structure of macromolecular crystals*. *J. Struct. Biol.* **117**, 124–137.

Malkin, A. J. & McPherson, A. (1994). *Light-scattering investigations of nucleation processes and kinetics of crystallization in macromolecular systems*. *Acta Cryst.* D**50**, 385–395.

Martins, P., Pëssoa, J., Sarkany, Z., Rocha, F. & Damas, A. (2008). *Rationalizing protein crystallization screenings through water equilibration theory and protein solubility data*. *Cryst. Growth Des.* **8**, 4233–4243.

Martins, P. M., Rocha, F. & Damas, A. M. (2008). *Understanding water equilibration fundamentals as a step for rational protein crystallization*. *PLoS One*, **3**, e1998.

Mikol, V., Hirsch, E. & Giegé, R. (1990). *Diagnostic of precipitant for biomacromolecule crystallization by quasi-elastic light-scattering*. *J. Mol. Biol.* **213**, 187–195.

Mikol, V., Rodeau, J.-L. & Giegé, R. (1989). *Changes of pH during biomacromolecule crystallization by vapor diffusion using ammonium sulfate as the precipitant*. *J. Appl. Cryst.* **22**, 155–161.

Mikol, V., Rodeau, J.-L. & Giegé, R. (1990). *Experimental determination of water equilibration rates in the hanging drop method of protein crystallization*. *Anal. Biochem.* **186**, 332–339.

Minezaki, Y., Niimura, N., Ataka, M. & Katsura, T. (1996). *Small angle neutron scattering from lysozyme solutions in unsaturated and supersaturated states (SANS from lysozyme solutions)*. *Biophys. Chem.* **58**, 355–363.

Mirzabekov, A. D., Rhodes, D., Finch, J. T., Klug, A. & Clark, B. F. (1972). *Crystallization of tRNAs as cetyltrimethylammonium salts*. *Nat. New Biol.* **237**, 27–28.

Mizianty, M. J. & Kurgan, L. (2009). *Meta prediction of protein crystallization propensity*. *Biochem. Biophys. Res. Commun.* **390**, 10–15.

Moreno, A., Saridakis, E. & Chayen, N. E. (2002). *Combination of oils and gels for enhancing the growth of protein crystals*. *J. Appl. Cryst.* **35**, 140–142.

Moreno, A. & Sazaki, G. (2004). *The use of a new ad hoc growth cell with parallel electrodes for the nucleation control of lysozyme*. *J. Cryst. Growth*, **264**, 438–444.

Moreno, A., Théobald-Dietrich, A., Lorber, B., Sauter, C. & Giegé, R. (2005). *Effects of macromolecular impurities and of crystallization method on the quality of eubacterial aspartyl-tRNA synthetase crystals*. *Acta Cryst.* D**61**, 789–792.

Moreno, A., Yokaichiya, F., Dimasi, E. & Stojanoff, V. (2009). *Growth and characterization of high-quality protein crystals for X-ray crystallography*. *Ann. N. Y. Acad. Sci.* **1161**, 429–436.

Nakada, T., Sazaki, G., Miyashita, S., Durbin, S. D. & Komatsu, H. (1999). *Direct AFM observations of impurity effects on a lysozyme crystal*. *J. Cryst. Growth*, **196**, 503–510.

Nanev, C. N. & Penkova, A. (2001). *Nucleation of lysozyme crystals under external electric and ultrasonic fields. J. Cryst. Growth*, **232**, 285–293.

Neal, B. L., Asthagiri, D., Velev, O. D., Lenhoff, A. M. & Kaler, E. W. (1999). *Why is the osmotic second virial coefficient related to protein crystallization? J. Cryst. Growth*, **196**, 377–387.

Neugebauer, J. M. (1990). *Detergents: an overview. Methods Enzymol.* **182**, 239–253.

Newman, J. (2004). *Novel buffer systems for macromolecular crystallization. Acta Cryst.* D**60**, 610–612.

Newman, J. (2005). *Expanding screening space through the use of alternative reservoirs in vapor-diffusion experiments. Acta Cryst.* D**61**, 490–493.

Newman, J., Pham, T. M. & Peat, T. S. (2008). *Phoenito experiments: combining the strengths of commercial crystallization automation. Acta Cryst.* F**64**, 991–996.

Newman, J., Xu, J. & Willis, M. C. (2007). *Initial evaluations of the reproducibility of vapor-diffusion crystallization. Acta Cryst.* D**63**, 826–832.

Ng, J., Lorber, B., Witz, J., Théobald-Dietrich, A., Kern, D. & Giegé, R. (1996). *The crystallization of macromolecules from precipitates: evidence for Ostwald ripening. J. Cryst. Growth*, **168**, 50–62.

Ng, J. D., Clark, P. J., Stevens, R. C. & Kuhn, P. (2008). *In situ X-ray analysis of protein crystals in low-birefringent and X-ray transmissive plastic microchannels. Acta Cryst.* D**64**, 189–197.

Ng, J. D., Gavira, J. A. & Garcia-Ruiz, J. M. (2003). *Protein crystallization by capillary counterdiffusion for applied crystallographic structure determination. J. Struct. Biol.* **142**, 218–231.

Ng, J. D., Kuznetsov, Y. G., Malkin, A. J., Keith, G., Giegé, R. & McPherson, A. (1997). *Visualization of RNA crystal growth by atomic force microscopy. Nucleic Acids Res.* **25**, 2582–2588.

Ng, J. D., Lorber, B., Giegé, R., Koszelak, S., Day, J., Greenwood, A. & McPherson, A. (1997). *Comparative analysis of thaumatin crystals grown on earth and in microgravity. Acta Cryst.* D**53**, 724–733.

Ng, J. D., Sauter, C., Lorber, B., Kirkland, N., Arnez, J. & Giegé, R. (2002). *Comparative analysis of space-grown and earth-grown crystals of an aminoacyl-tRNA synthetase: space-grown crystals are more useful for structural determination. Acta Cryst.* D**58**, 645–652.

Niesen, F. H., Koch, A., Lenski, U., Harttig, U., Roske, Y., Heinemann, U. & Hofmann, K. P. (2008). *An approach to quality management in structural biology: biophysical selection of proteins for successful crystallization. J. Struct. Biol.* **162**, 451–459.

Nollert, P. (2004). *Lipidic cubic phases as matrices for membrane protein crystallization. Methods*, **34**, 348–353.

Ostermeier, C., Iwata, S., Ludwig, B. & Michel, H. (1995). *F-V fragment mediated crystallization of the membrane-protein bacterial cytochrome-C-oxidase. Nat. Struct. Biol.* **2**, 842–846.

Otálora, F., Garcia-Ruiz, J. M., Gavira, J. A. & Capelle, B. (1999). *Topography and high resolution diffraction studies in tetragonal lysozyme. J. Cryst. Growth*, **196**, 546–558.

Otálora, F., Gavira, J. A., Ng, J. D. & Garcia-Ruiz, J. M. (2009). *Counterdiffusion methods applied to protein crystallization. Prog. Biophys. Mol. Biol.* **101**, 26–37.

Papanikolau, Y. & Kokkinidis, M. (1997). *Solubility, crystallization, and chromatographic properties of macromolecules strongly depend on substances that reduce the ionic strength of the solution. Protein Eng.* **10**, 847–850.

Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Tapping the Protein Data Bank for crystallization information. Acta Cryst.* D**61**, 1662–1669.

Price, W. N., Chen, Y., Handelman, S. K., Neely, H., Manor, P., Karlin, R., Nair, R., Liu, J., Baran, M., Everett, J., Tong, S. N., Forouhar, F., Swaminathan, S. S., Acton, T., Xiao, R., Luft, J. R., Lauricella, A., DeTitta, G. T., Rost, B., Montelione, G. T. & Hunt, J. F. (2009). *Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data. Nat. Biotechnol.* **27**, 51–57.

Provost, K. & Robert, M.-C. (1995). *Crystal growth of lysozymes in media contaminated by parent molecules: influence of gelled media. J. Cryst. Growth*, **156**, 112–120.

Pusey, M. L., Liu, Z. J., Tempel, W., Praissman, J., Lin, D., Wang, B. C., Gavira, J. A. & Ng, J. D. (2005). *Life in the fast lane for protein crystallization and X-ray crystallography. Prog. Biophys. Mol. Biol.* **88**, 359–386.

Pusey, M. L., Paley, M. S., Turner, M. B. & Rogers, R. D. (2007). *Protein crystallization using room temperature ionic liquids. Cryst. Growth Des.* **7**, 787–793.

Qi, J., Wakayama, N. I. & Ataka, M. (2001). *Magnetic suppression of convection in protein crystal growth processes. J. Cryst. Growth*, **232**, 132–137.

Ray, W. J. & Puvathingal, J. M. (1985). *A simple procedure for removing contaminating aldehydes and peroxides from aqueous solutions of polyethylene glycols and of nonionic detergents that are based on the polyoxyethylene linkage. Anal. Biochem.* **146**, 307–312.

Ray, W. J. J. & Bracker, C. E. (1986). *Polyethylene glycol: catalytic effect on the crystallization of phosphoglucomutase at high salt concentration. J. Cryst. Growth*, **76**, 562–576.

Reid, B. R., Koch, G. L. E., Boulanger, Y., Hartley, B. S. & Blow, D. (1973). *Crystallization and preliminary X-ray diffraction studies on tyrosyl transfer RNA synthetase from Bacillus stearothermophilus. J. Mol. Biol.* **80**, 199–201.

Rhim, W.-K. & Chung, S. K. (1990). *Isolation of crystallizing droplets by electrostatic levitation. Methods*, **1**, 118–127.

Richard, B., Bonneté, F., Dym, O. & Zaccaï, G. (1995). *The MPD-NaCl-H_2O system for the crystallization of halophilic proteins*. In *Archaea, a Laboratory Manual*, pp. 149–154. Cold Spring Harbor Laboratory Press.

Riès-Kautt, M. & Ducruix, A. (1991). *Crystallization of basic proteins by ion pairing. J. Cryst. Growth*, **110**, 20–25.

Robert, M.-C., Bernard, Y. & Lefaucheux, F. (1994). *Study of nucleation-related phenomena in lysozyme solutions. Application to gel growth. Acta Cryst.* D**50**, 496–503.

Robert, M.-C. & Lefaucheux, F. (1988). *Crystal growth in gels: principle and applications. J. Cryst. Growth*, **90**, 358–367.

Rosenberger, F. (1996). *Protein crystallization. J. Cryst. Growth*, **166**, 40–54.

Rosenberger, F., Vekilov, P. G., Muschol, M. & Thomas, B. R. (1996). *Nucleation and crystallization of globular proteins – what do we know and what is missing. J. Cryst. Growth*, **168**, 1–27.

Salemme, F. R. (1972). *A free interface diffusion technique for crystallization of proteins for X-ray crystallography. Arch. Biochem. Biophys.* **151**, 533–540.

Saridakis, E. & Chayen, N. E. (2009). *Towards a 'universal' nucleant for protein crystallization. Trends Biotechnol.* **27**, 99–106.

Sato, T., Yamada, Y., Saijo, S., Hori, T., Hirose, S., Tanaka, N., Sazaki, G., Nakajima, K., Igarashi, N., Tanaka, M. & Matsuura, Y. (2000). *Enhancement in the perfection of orthorhombic lysozyme crystals grown in a high magnetic field (10 T). Acta Cryst.* D**56**, 1079–1083.

Sauter, C., Balg, C., Moreno, A., Dhouib, K., Théobald-Dietrich, A., Chênevert, R., Giegé, R. & Lorber, B. (2009). *Agarose gel facilitates enzyme crystal soaking with a ligand analog. J. Appl. Cryst.* **42**, 279–283.

Sauter, C., Dhouib, K. & Lorber, B. (2007). *From macrofluidics to microfluidics for the crystallization of biological macromolecules. Cryst. Growth Des.* **7**, 2247–2250.

Sauter, C., Lorber, B. & Giegé, R. (2002). *Towards atomic resolution with crystals grown in gel: the case of thaumatin seen at room temperature. Proteins*, **48**, 146–150.

Sauter, C., Lorber, B., Kern, D., Cavarelli, J., Moras, D. & Giegé, R. (1999). *Crystallogenesis studies on yeast aspartyl-tRNA synthetase: use of phase diagram to improve crystal quality. Acta Cryst.* D**55**, 149–156.

Sauter, C., Ng, J. D., Lorber, B., Keith, G., Brion, P., Hosseini, M. W., Lehn, J.-M. & Giegé, R. (1999). *Additives for the crystallization of proteins and nucleic acids. J. Cryst. Growth*, **196**, 365–376.

Sazaki, G., Moreno, A. & Nakajima, K. (2004). *Novel coupling effects of the magnetic and electric fields on protein crystallization. J. Cryst. Growth*, **262**, 499–502.

Sazaki, G., Okada, M., Matsui, T., Watanabe, T., Higuchi, H., Tsukamoto, K. & Nakajima, K. (2008). *Single-molecule visualization of diffusion at the solution-crystal interface. Cryst. Growth Des.* **8**, 2024–2031.

Sazaki, G., Tsukamoto, K., Yai, S., Okada, M. & Nakajima, K. (2005). *In situ observation of dislocations in protein crystals during growth by advanced optical microscopy. Cryst. Growth Des.* **5**, 1729–1735.

Sazaki, G., Yoshida, E., Komatsu, H., Nakada, T., Miyashita, S. & Watanabe, K. (1997). *Effects of a magnetic field on the nucleation and growth of protein crystals. J. Cryst. Growth*, **173**, 231–234.

Sennhauser, G. & Grütter, M. G. (2008). *Chaperone-assisted crystallography with DARPins. Structure*, **16**, 1443–1453.

Shaw, N., Cheng, C., Tempel, W., Chang, J., Ng, J., Wang, X. Y., Perrett, S., Rose, J., Rao, Z., Wang, B. C. & Liu, Z. J. (2007). *(NZ)CH···O contacts assist crystallization of a ParB-like nuclease*. *BMC Struct. Biol.* **7**, 46.

Shim, J. U., Cristobal, G., Link, D. R., Thorsen, T., Jia, Y., Piattelli, K. & Fraden, S. (2007). *Control and measurement of the phase behavior of aqueous solutions using microfluidics*. *J. Am. Chem. Soc.* **129**, 8825–8835.

Shu, Z.-Y., Gong, H.-Y. & Bi, R.-C. (1998). *In situ measurements and dynamic control of the evaporation rate in vapor diffusion crystallization of proteins*. *J. Cryst. Growth*, **192**, 282–289.

Snell, E. H., Helliwell, J. R., Boggon, T. J., Lautenschlager, P. & Potthast, L. (1996). *Lysozyme crystal growth kinetics monitored using a Mach–Zehnder interferometer*. *Acta Cryst.* D**52**, 529–533.

Snell, E. H., Weisgerber, S., Helliwell, J. R., Weckert, E., Hölzer, K. & Schroer, K. (1995). *Improvements in lysozyme protein crystal perfection through microgravity growth*. *Acta Cryst.* D**51**, 1099–1102.

Song, H., Chen, D. L. & Ismagilov, R. F. (2006). *Reactions in droplets in microfluidic channels*. *Angew. Chem. Int. Ed. Engl.* **45**, 7336–7356.

Sousa, R. (1995). *Use of glycerol, polyols and other protein structure stabilizing agents in protein crystallization*. *Acta Cryst.* D**51**, 271–277.

Stojanoff, V., Siddons, D. P., Monaco, L. A., Vekilov, P. & Rosenberger, F. (1997). *X-ray topography of tetragonal lysozyme grown by the temperature-controlled technique*. *Acta Cryst.* D**53**, 588–595.

Stojanoff, V., Snell, E. F., Siddons, D. P. & Helliwell, J. R. (1996). *An old technique with a new application: X-ray topography of protein crystals*. *Synchrotron Radiat. News*, **9**, 25–26.

Stolyarova, S., Saridakis, E., Chayen, N. E. & Nemirovsky, Y. (2006). *A model for enhanced nucleation of protein crystals on a fractal porous substrate*. *Biophys. J.* **91**, 3857–3863.

Strickland, C. L., Puchalsky, R., Savvides, S. N. & Karplus, P. A. (1995). *Overexpression of Crithidia fasciculata trypanothione reductase and crystallization using a novel geometry*. *Acta Cryst.* D**51**, 337–341.

Stura, E. A., Charbonnier, J.-B. & Taussig, M. J. (1999). *Epitaxial jumps*. *J. Cryst. Growth*, **196**, 250–260.

Sugahara, M., Kageyama, Y. & Kunishima, N. (2009). *A hands-free capillary system for protein crystallography from crystallization to data collection*. *J. Appl. Cryst.* **42**, 129–133.

Sugiyama, S., Tanabe, K., Hirose, M., Kitatani, T., Hasenaka, H., Takahashi, Y., Adahi, H., Takano, K., Murakami, S., Mori, Y., Inoue, T. & Matsumura, S. (2009). *Protein crystallization in agarose gel with high strength: developing an automated system for protein crystallographic processes*. *Jpn. J. Appl. Phys.* **48**, 075502.

Suzuki, Y., Sazaki, G., Miyashita, S., Sawada, T., Tamura, K. & Komatsu, H. (2002). *Protein crystallization under high pressure*. *Biochim. Biophys. Acta*, **1595**, 345–356.

Taleb, M., Didierjean, C., Jelsch, C., Mangeot, J.-P., Capelle, B. & Aubry, A. (1999). *Crystallization of proteins under an external electric field*. *J. Cryst. Growth*, **200**, 575–582.

Thakur, A. S., Newman, J., Martin, J. L. & Kobe, B. (2008). *Increasing protein crystallization screening success with heterogeneous nucleating agents*. *Methods Mol. Biol.* **426**, 403–409.

Thaller, C., Eichele, G., Weaver, L. H., Wilson, E., Karlsson, R. & Jansonius, J. N. (1985). *Seed enlargement and repeated seeding*. *Methods Enzymol.* **114**, 132–135.

Thomas, B. R., Vekilov, P. G. & Rosenberger, F. (1998). *Effects of microheterogeneity in hen egg-white lysozyme crystallization*. *Acta Cryst.* D**54**, 226–236.

Thomas, D. H., Rob, A. & Rice, D. W. (1989). *A novel dialysis procedure for the crystallization of proteins*. *Protein Eng.* **2**, 489–491.

Tung, M. & Gallagher, D. T. (2009). *The Biomolecular Crystallization Database Version 4: expanded content and new features*. *Acta Cryst.* D**65**, 18–23.

Van Driessche, A. E. S., Otalora, F., Gavira, J. A. & Sazaki, G. (2008). *Is agarose an impurity or an impurity filter? In situ observation of the joint gel/impurity effect on protein crystal growth kinetics*. *Cryst. Growth Des.* **8**, 3623–3629.

Van Driessche, A. E. S., Otalora, F., Sazaki, G., Sleutel, M., Tsukamoto, K. & Gavira, J. A. (2008). *Comparison of different experimental techniques for the measurement of crystal growth kinetics*. *Cryst. Growth Des.* **8**, 4316–4323.

Vaney, M. C., Maignan, M., Riès-Kautt, M. & Ducruix, A. (1996). *High-resolution structure (1.33 Å) of a HEW lysozyme tetragonal crystal grown in the APCF apparatus. Data and structural comparison with a crystal grown under microgravity from Spacehab-01 mission*. *Acta Cryst.* D**52**, 505–517.

Veesler, S., Furuta, K., Horiuchi, H., Hiratsuka, H., Ferte, N. & Okutsu, T. (2006). *Crystals from light: photochemically induced nucleation of hen egg-white lysozyme*. *Cryst. Growth Des.* **6**, 1631–1635.

Vekilov, P. G., Ataka, M. & Katsura, T. (1993). *Laser Michelson interferometry investigation of protein crystal growth*. *J. Cryst. Growth*, **130**, 317–320.

Vekilov, P. G. & Rosenberger, F. (1996). *Dependence of lysozyme growth kinetics on step sources and impurities*. *J. Cryst. Growth*, **158**, 540–551.

Vekilov, P. G. & Rosenberger, F. (1998). *Protein crystal growth under forced solution flow: experimental setup and general response of lysozyme*. *J. Cryst. Growth*, **186**, 251–261.

Vergara, A., Lorber, B., Sauter, C., Giegé, R. & Zagari, A. (2005). *Lessons from crystals grown in the Advanced Protein Crystallization Facility for conventional crystallization applied to structural biology*. *Biophys. Chem.* **118**, 102–112.

Vergara, A., Lorber, B., Zagari, A. & Giegé, R. (2003). *Physical aspects of protein crystal growth investigated with the Advanced Protein Crystallization Facility in reduced-gravity environments*. *Acta Cryst.* D**59**, 2–15.

Vidal, O., Robert, M.-C., Arnoux, B. & Capelle, B. (1999). *Crystalline quality of lysozyme crystals grown in agarose and silica gels studied by X-ray diffraction techniques*. *J. Cryst. Growth*, **196**, 559–571.

Vidal, O., Robert, M.-C. & Boué, F. (1998). *Gel growth of lysozyme crystals studied by small angle neutron scattering: case of agarose gel, a nucleation promotor*. *J. Cryst. Growth*, **192**, 257–270.

Viola, R., Carman, P., Walsh, J., Miller, E., Benning, M., Frankel, D., McPherson, A., Cudney, B. & Rupp, B. (2007). *Operator-assisted harvesting of protein crystals using a universal micromanipulation robot*. *J. Appl. Cryst.* **40**, 539–545.

Warke, A. & Momany, C. (2007). *Addressing the protein crystallization bottleneck by cocrystallization*. *Cryst. Growth Des.* **7**, 2219–2225.

Washabaugh, M. W. & Collins, K. D. (1986). *The systematic characterization by aqueous column chromatography of solutes which affect protein stability*. *J. Biol. Chem.* **261**, 2477–2485.

Weber, B. H. & Goodkin, P. E. (1970). *A modified microdiffusion procedure for the growth of single protein crystals by concentration-gradient equilibrium dialysis*. *Arch. Biochem. Biophys.* **141**, 489–498.

Wernimont, A. & Edwards, A. (2009). *In situ proteolysis to generate crystals for structure determination: an update*. *PLoS One*, **4**, e5094.

Wiener, M. (2004). *A pedestrian guide to membrane protein crystallization*. *Methods*, **34**, 364–372.

Wilson, W. W. (2003). *Light scattering as a diagnostic for protein crystal growth – a practical approach*. *J. Struct. Biol.* **142**, 56–65.

Yadav, M. K., Gerdts, C. J., Sanishvili, R., Smith, W. W., Roach, L. S., Ismagilov, R. F., Kuhn, P. & Stevens, R. C. (2005). *In situ data collection and structure refinement from microcapillary protein crystallization*. *J. Appl. Cryst.* **38**, 900–905.

Ye, J. D., Tereshko, V., Frederiksen, J. K., Koide, A., Fellouse, F. A., Sidhu, S. S., Koide, S., Kossiakoff, A. A. & Piccirilli, J. A. (2008). *Synthetic antibodies for specific recognition and crystallization of structured RNA*. *Proc. Natl Acad. Sci. USA*, **105**, 82–87.

Yonath, A., Müssig, J. & Wittmann, H. G. (1982). *Parameters for crystal growth of ribosomal subunits*. *J. Cell. Biochem.* **19**, 145–155.

Yoshikawa, H. Y., Murai, R., Sugiyama, S., Sazaki, G., Kitatani, T., Takahashi, Y., Adachi, H., Matsumura, H., Murakami, S., Inoue, T., Takano, K. & Mori, Y. (2009). *Femtosecond laser-induced nucleation of protein in agarose gel*. *J. Cryst. Growth*, **311**, 956–959.

Zeppenzauer, M. (1971). *Formation of large crystals*. *Methods Enzymol.* **22**, 235–266.

Zheng, B., Roach, L. S. & Ismagilov, R. F. (2003). *Screening of protein crystallization conditions on a microfluidic chip using nanoliter-size droplets*. *J. Am. Chem. Soc.* **125**, 11170–11171.

Zhu, D.-W., Lorber, B., Sauter, C., Ng, J. D., Bénas, P., Le Grimellec, C. & Giegé, R. (2001). *Growth kinetics, diffraction properties and effect of agarose on the stability of a novel crystal form of Thermus thermophilus aspartyl-tRNA synthetase-1*. *Acta Cryst.* D**57**, 552–558.

Zulauf, M. (1990). *Detergent phenomena in membrane protein crytallization*. In *Crystallization of Membrane Proteins*, edited by H. Michel, ch. 2, pp. 53–72. Boca Raton: CRC Press.

International Tables for Crystallography (2012). Vol. F, Chapter 4.2, pp. 122–128.

Chapter 4.2. Crystallization of membrane proteins

H. Michel

4.2.1. Introduction

At the time of writing, the Protein Data Bank contains more than 8000 entries of protein structures. These belong to roughly 1200 sequence-unrelated protein families, which can be classified into ~350 different folds (Gerstein, 1998). However, all known membrane-protein structures belong to one of a dozen membrane-protein families. Membrane proteins for which the structures have been determined are bacterial photosynthetic reaction centres, porins and other β-strand barrel-forming proteins from the outer membrane of Gram-negative bacteria, bacterial light-harvesting complexes, bacterial and mitochondrial cytochrome *c* oxidase, the cytochrome bc_1 complex from mammalian mitochondria, two cyclooxygenases, squalene cyclase, and two bacterial channel proteins. These structures have been determined by X-ray crystallography. In addition, the structures of two membrane proteins, namely that of bacteriorhodopsin and that of the plant light-harvesting complex II, have been solved by electron crystallography (see Chapter 19.6). Table 4.2.1.1 provides a list of the membrane proteins with known structures. It also contains the key references for the structure descriptions and the crystallization conditions.

It is known from genome sequencing projects that 20–35% of all proteins contain at least one transmembrane segment (Gerstein, 1998), as deduced from the occurrence of stretches of hydrophobic amino acids that are long enough to span the membrane in a helical manner. These numbers may be an underestimate, because β-strand-rich membrane proteins like the porins, or membrane proteins which are only inserted into the membrane, but do not span it ('monotopic membrane proteins'), like cyclooxygenases (prostaglandin-H synthases), cannot be recognized as being membrane proteins by inspecting their amino-acid sequences.

Why do we know so few membrane-protein structures? The first reason is the lack of sufficient amounts of biochemically well characterized, homogeneous and stable membrane-protein preparations. This is especially true for eukaryotic receptors and transporters (we do not know the structures of any of these proteins). For these, a major problem is the lack of efficient expression systems for heterologous membrane-protein production. It is therefore not surprising that most membrane proteins with known structure are either involved in photosynthesis or bioenergetics (they are relatively abundant), or originate from bacterial outer membranes (they are exceptionally stable and can be overproduced). Second, membrane proteins are integrated into membranes. They have two polar surface regions on opposite sides (where they are in contact with the aqueous phases and the polar head groups of the membrane lipids) which are separated by a hydrophobic belt. The latter is in contact with the alkyl chains of the lipids. As a result of this amphipathic nature of their surface, membrane proteins are not soluble in either aqueous or organic solvents. To isolate membrane proteins one first has to prepare the membranes, and then solubilize the membrane proteins by adding an excess of detergent. Detergents consist of a polar or charged head group and a hydrophobic tail. Above a certain concentration, the so-called *critical micellar concentration* (CMC), detergents form *micelles* by association of their hydrophobic tails. These micelles take up lipids. Detergents also bind to the hydrophobic surface of membrane proteins with their hydrophobic tails and form a ring-like detergent micelle surrounding the membrane protein, thus shielding the hydrophobic belt-like surface of the membrane protein from contact with water. This is the reason for their ability to solubilize membrane proteins, although with detergents with large polar head groups it is sometimes difficult to achieve a rapid and complete solubilization. The solubilizate, consisting of these mixed protein–detergent complexes as well as lipid-containing and pure detergent micelles, is then subjected to similar purification procedures as are soluble proteins. Of course, the presence of detergents complicates the purification procedures. The choice of the detergent is critical. The detergent micelles have to replace, and to mimic, the lipid bilayer as perfectly as possible, in order to maintain the stability and activity of the solubilized membrane protein. The solubilization of membrane proteins has been reviewed by Hjelmeland (1990) and the general properties of the detergents used has been reviewed by Neugebauer (1990).

4.2.2. Principles of membrane-protein crystallization

There are two principal types of membrane-protein crystals (Michel, 1983). First, one can think of forming two-dimensional crystals in the planes of the membrane and stacking these two-dimensional crystals in an ordered way with respect to up and down orientation, rotation and translation. This membrane-protein crystal type ('type I') is attractive, because it contains the membrane proteins in their native environment. It should even be possible to study lipid–protein interactions. Crystals of bacteriorhodopsin of this type have been obtained either upon slow removal of the detergent by dialysis at high ionic strength (Henderson & Shotton, 1980), or by a novel approach using lipidic bicontinuous cubic phases (Landau & Rosenbusch, 1996; Pebay-Peyroula *et al.*, 1997; see also below). Alternatively, one can try to crystallize the membrane protein with the detergents still bound in a micellar manner. These crystals are held together *via* polar interactions between the polar surfaces of the membrane proteins. The detergent plays a more passive, but still critical, role. Such 'type II' crystals look very much like crystals of soluble globular proteins. The same crystallization methods and equipment as for soluble globular proteins (see Chapter 4.1) can be used. However, the use of hanging drops is sometimes difficult, because the presence of detergents leads to a lower surface tension of the protein solution. Intermediate forms between type I and type II crystals are feasible, *e.g.* by fusion of detergent micelles.

The use of detergent concentrations just above the CMC of the respective detergent is recommended in order to prevent complications caused by pure detergent micelles. Unfortunately, the CMC is not constant. Normally, the CMC provided by the vendor has been determined in water at room temperature. A compilation of potentially useful detergents, their CMCs and their molecular weights is presented in Table 4.2.2.1. The CMC is

Copyright © 2012 International Union of Crystallography

Table 4.2.1.1

Compilation of membrane proteins with known structures, including crystallization conditions and key references for the structure determinations

This table is continuously updated and can be inspected at http://www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html. The membrane proteins listed are divided into polytopic membrane proteins from inner membranes of bacteria and mitochondria (*a*), membrane proteins from the outer membrane of Gram-negative bacteria (*b*) and monotopic membrane proteins [(*c*); these are proteins that are only inserted into the membrane, but do not span it]. Within parts (*a*), (*b*) and (*c*) the membrane proteins are listed in chronological order of structure determination.

(*a*) Polytopic membrane proteins from inner membranes of bacteria and mitochondria.

Membrane protein	Crystallization conditions (detergent/additive/precipitating agent)	Key references (and pdb reference code, if available)
Photosynthetic reaction centre		
from *Rhodopseudomonas viridis*	*N*,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol/ammonium sulfate	[1], [2] (1PRC), [3], [4] (2PRC, 3PRC, 4PRC, 5PRC, 6PRC, 7PRC)
from *Rhodobacter sphaeroides*	*N*,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol/polyethylene glycol 4000	[5] (4RCR)
	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[6] (2RCR)
	N,*N*-Dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol, dioxane/potassium phosphate	[7] (1PCR)
	Octyl-β-D-glucopyranoside/benzamidine, heptane-1,2,3-triol/polyethylene glycol 4000	[8] (1AIG, 1AIJ)
Bacteriorhodopsin		
from *Halobacterium salinarium*	(Electron crystallography using naturally occuring two-dimensional crystals)	[9] (1BRD), [10] (2BRD), [11] (1AT9)
	(Type I crystal grown in lipidic cubic phases)	[12] (1AP9), [13] (1BRX)
	Octyl-β-D-glucopyranoside/benzamidine/sodium phosphate (epitaxic growth on benzamidine crystals)	[14] (1BRR)
Light-harvesting complex II		
from pea chloroplasts	(Electron crystallography of two-dimensional crystals prepared from Triton X100 solubilized material)	[15]
Light-harvesting complex 2		
from *Rhodopseudomonas acidophila*	Octyl-β-D-glucopyranoside/benzamidine/phosphate	[16] (1KZU)
from *Rhodospirillum molischianum*	*N*,*N*-dimethylundecylamine-*N*-oxide/heptane-1,2,3-triol/ammonium sulfate	[17] (1LGH)
Cytochrome *c* oxidase		
from *Paracoccus denitrificans*,		
four-subunit enzyme complexed with antibody Fv fragment	Dodecyl-β-D-maltoside/polyethylene glycol monomethylether 2000	[18]
two-subunit enzyme complexed with antibody Fv fragment	Undecyl-β-D-maltoside/polyethylene glycol monomethylether 2000	[19] (1AR1)
from bovine heart mitochondria	Decyl-β-D-maltoside with some residual cholate/polyethylene glycol 4000	[20], [21] (1OCC), [22] (2OCC, 1OCR)
Cytochrome bc_1 complex		
from bovine heart mitochondria	Decanoyl-*N*-methylglucamide or diheptanoyl phosphatidyl choline/polyethylene glycol 4000	[23] (1QRC), [24]
	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[25]
	Pure dodecyl-β-D-maltoside or mixture with methyl-6-*O*-(*N*-heptylcarbamoyl)-α-D-glucopyranoside/polyethylene glycol 4000	[26]
from chicken heart mitochondria	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[25] (1BCC, 3BCC)
Potassium channel		
from *Streptomyces lividans*	*N*,*N*-Dimethyldodecylamine/polyethylene glycol 400	[27] (1BL8)
Mechanosensitive ion channel		
from *Mycobacterium tuberculosis*	Dodecyl-β-D-maltoside/triethylene glycol	[28]

generally lower at high ionic strength and at high temperatures. The presence of glycerol and similar compounds, as well as that of chaotropic agents (Midura & Yanagishita, 1995), also influences (decreases) the CMC.

4.2.3. General properties of detergents relevant to membrane-protein crystallization

The presence of detergents sometimes causes problems. The monomeric detergent itself can crystallize, *e.g.* dodecyl-β-D-maltoside at 4 °C in the presence of polyethylene glycol. The detergent crystals might be mistaken for protein crystals. Detergent micelles possess attractive interactions (see Zulauf, 1991). Upon addition of salts or polyethylene glycol, or upon temperature changes, a phase separation may be observed: owing to an increase in these attractive interactions, the detergent micelles 'precipitate', forming a viscous detergent-rich phase and a detergent-depleted aqueous phase. The membrane proteins are found exclusively in the viscous phase and crystals – if formed – are difficult to handle. Some detergents, *e.g.* those with polyoxyethylene head groups, undergo a phase separation at higher temperatures. This phenomenon has been used to separate

Table 4.2.1.1 (continued)

(*b*) Membrane proteins from the outer membrane of Gram-negative bacteria and related proteins.

Membrane protein	Crystallization conditions (detergent/additive/ precipitating agent)	Key references (and pdb reference code, if available)
16-Stranded porins		
from *Rhodobacter capsulatus*	Octytetraoxyethylene/polyethylene glycol 600	[29] (2POR)
OmpF and PhoE from *Escherichia coli*	Mixture of *n*-octyl-2-hydroxyethylsulfoxide and octylpolyoxyethylene; or *N*,*N*-dimethyldecylamine-*N*-oxide/polyethylene glycol 2000	[30] (1OPF, 1PHO), [31]
from *Rhodopseudomonas blastica*	Octyltetraoxyethylene/heptane-1,2,3-triol/ polyethylene glycol 600	[32] (1PRN)
from *Paracoccus denitrificans*	Octyl-β-D-glucoside/polyethylene glycol 600	[33]
18-Stranded porins		
maltoporin from *Escherichia coli*	Mixture of decyl-β-D-maltoside and dodecylnonaoxyethylene/polyethylene glycol 2000	[34] (1MAL)
maltoporin from *Salmonella typhimurium*	Mixture of octyltetraoxyethylene and *N*,*N*-dimethylhexylamine-*N*-oxide/polyethylene glycol 1500	[35] (1MPR, 2MPR)
sucrose-specific ScrY porin from *Salmonella typhimurium*	Mixture of octyl-β-D-glucopyranoside and *N*,*N*-dimethylhexylamine-*N*-oxide/polyethylene glycol 2000	[36] (1AOS, 1AOT)
α-Haemolysin		
from *Staphylococcus aureus*	Octyl-β-D-glucopyranoside/ammonium sulfate, polyethylene glycol monomethylether 5000	[37] (7AHL)
Eight-stranded β-barrel membrane anchor		
OmpA fragment from *Escherichia coli*	Not yet available	[38] (1BXW)
22-Stranded receptors		
FhuA from *Escherichia coli*	*N*,*N*-Dimethyldecylamine-*N*-oxide/inositol/ polyethylene glycol monomethylether 2000	[39] (1FCP, 2FCP)
	n-Octyl-2-hydroxyethylsulfoxide/polyethylene glycol 2000	[40] (1BY3, 1BY5)
ferric enterobacterin receptor (FepA) from *Escherichia coli*	*N*,*N*-dimethyldodecylamine-*N*-oxide/heptane-1,2,3-triol/polyethylene glycol 1000	[41] (1FEP)

(*c*) Proteins inserted into, but not crossing the membrane ('monotopic membrane proteins').

Membrane protein	Crystallization conditions (detergent/additive/ precipitating agent)	Key references (and pdb reference code, if available)
Prostaglandin H_2 synthase 1		
(cyclooxygenase 1) from sheep	Octyl-β-D-glucopyranoside/polyethylene glycol 4000	[42] (1PRH)
Cyclooxygenase 2		
from mouse	Octyl-β-D-glucopyranoside/polyethylene glycol monomethylether 550	[43] (1CX2, 3PGH, 4COX, 5COX, 6COX)
from man	Octylpentaoxyethylene/polyethylene glycol 4000	[44]
Squalene cyclase		
from *Alicyclobacillus acidocaldarius*	Octyltetraoxyethylene/sodium citrate	[45] (1SQC)

References: [1] Diesenhofer *et al*. (1985); [2] Diesenhofer *et al*. (1995); [3] Lancaster & Michel (1997); [4] Lancaster & Michel (1999); [5] Allen *et al*. (1987); [6] Chang *et al*. (1991); [7] Ermler *et al*. (1994); [8] Stowell *et al*. (1997); [9] Henderson *et al*. (1990); [10] Grigorieff *et al*. (1996); [11] Kimura *et al*. (1997); [12] Pebay-Peyroula *et al*. (1997); [13] Luecke *et al*. (1998); [14] Essen *et al*. (1998); [15] Kühlbrandt *et al*. (1994); [16] McDermott *et al*. (1995); [17] Koepke *et al*. (1996); [18] Iwata *et al*. (1995); [19] Ostermeier *et al*. (1997); [20] Tsukihara *et al*. (1995); [21] Tsukihara *et al*. (1996); [22] Yoshikawa *et al*. (1998); [23] Xia *et al*. (1997); [24] Kim *et al*. (1998); [25] Zhang *et al*. (1998); [26] Iwata *et al*. (1998); [27] Doyle *et al*. (1998); [28] Chang *et al*. (1998); [29] Weiss *et al*. (1991); [30] Cowan *et al*. (1992); [31] Cowan *et al*. (1995); [32] Kreusch *et al*. (1994); [33] Hirsch *et al*. (1997); [34] Schirmer *et al*. (1995); [35] Meyer *et al*. (1997); [36] Forst *et al*. (1998); [37] Song *et al*. (1996); [38] Pautsch & Schulz (1998); [39] Ferguson *et al*. (1998); [40] Locher *et al*. (1998); [41] Buchanan *et al*. (1999); [42] Picot *et al*. (1994); [43] Kurumbail *et al*. (1996); [44] Luong *et al*. (1996); [45] Wendt *et al*. (1997).

solubilized membrane proteins, which are found in the detergent-rich phase, from the water-soluble proteins. The latter are concentrated in the detergent-depleted phase (Bordier, 1981). Other detergents, *e.g.* octyl-β-D-glucopyranoside, show this phase separation at lower temperatures. Therefore, if phase separation causes problems, a change of the crystallization temperature may help.

The polar head groups of the detergents influence their usage in many ways. One would like to have a small polar head group, because the head group 'covers' the part of the protein's polar surface that is adjacent to the hydrophobic surface belt. The bigger the head group the more of the polar surface is covered and unavailable for the polar interactions needed to form the crystal lattice. Unfortunately, detergents with small polar head groups are rather denaturing. Detergents with charged head groups cannot be used, but detergents with zwitterionic head groups, *e.g.* sulfobetaines, can be tried with more stable proteins. The head group of a very successful detergent, *N*,*N*-dimethyldodecylamine-*N*-oxide, is of zwitterionic nature. I estimate that it can only be used with about 20% of all membrane proteins. Detergents with sugar residues as head groups have been used successfully. Octyl-β-D-glucopyranoside also tends to be denaturing. The lifetime of many membrane proteins can be prolonged by a factor of three by the use of nonyl-β-D-glucopyranoside instead of the shorter homologue. Such behaviour is observed within each series of homologous detergents; an increase in the alkyl chain by one methylene group leads to an increase in stability by a factor of three, an increase by two methylene groups leads to an increase in stability by a factor of about ten. Unfortunately, decyl-β-D-glucopyranoside is too insoluble to be used as detergent. For less stable membrane proteins, alkylmaltoside detergents or alkanoylsucrose deter-

Table 4.2.2.1

Potentially useful detergents for membrane-protein crystallizations with molecular weights and CMCs [in water, from Michel (1991) or as provided by the vendor]

The lengths of the alkyl or alkanoyl side chains are given as C_6 to C_{16}.

Detergent	Molecular weight	CMC (m*M*)
Alkyl dihydroxypropyl sulfoxide		
C_8	238	20.6
N,*N*-Dimethylalkylamine-*N*-oxides		
C_8	173	162
C_9	187	50
C_{10}	201	20
C_{12}	229	1–2
n-Dodecyl-*N*,*N*-dimethylglycine (zwitterionic)	271	1.5
N-Alkyl-β-D-glucopyranosides		
C_6	264	250
C_7	278	79
C_8	292	30
C_9	306	6.5
C_{10}	320	2.6
n-Alkanoyl-*N*-hydroxyethylglucamides ('HEGA-*n*')		
C_8	351	109
C_9	365	39
C_{10}	379	7.0
C_{11}	393	1.4
Alkyl hydroxyethyl sulfoxide		
C_8	222	15.8
n-Alkyl-β-D-maltosides		
C_6	426	210
C_8	454	19.5
C_9	468	6
C_{10}	483	1.8
C_{11}	497	0.6
C_{12}	511	0.17
C_{13}	525	0.033
C_{14}	539	0.01
C_{16}	567	0.006
cyclohexyl-C_1	438	340
cyclohexyl-C_2	452	120
cyclohexyl-C_3	466	34.5
cyclohexyl-C_4	480	7.6
cyclohexyl-C_5	494	2.4
cyclohexyl-C_6	508	0.56
cyclohexyl-C_7	522	0.19
n-Alkanoyl-*N*-methylglucamides ('MEGA-*n*')		
C_8	321	79
C_9	335	25
C_{10}	349	6
Methyl-6-*O*-(*N*-heptylcarbamoyl)-α-D-glucopyranoside ('HECAMEG')	335	19.5
n-Alkylphosphocholines (zwitterionic)		
C_8	295	114
C_9	309	39.5
C_{10}	323	11
C_{12}	315	1.5
C_{14}	379	0.12
C_{16}	407	0.013
Polyoxyethylene monoalkylethers (C_nE_m)		
C_8E_4	306	7.9
C_8E_5	350	7.1
$C_{10}E_6$	422	0.9
$C_{12}E_8$	538	0.071
n-Alkanoylsucrose		
C_{10}	497	2.5
C_{12}	525	0.3
n-Alkyl-β-D-thioglucopyranosides		
C_7	294	29
C_8	308	9
C_9	322	2.9
C_{10}	336	0.9
n-Alkyl-β-D-thiomaltopyranosides		
C_8	471	8.5
C_9	485	3.2
C_{10}	499	0.9
C_{11}	513	0.21
C_{12}	527	0.05

gents have to be tried. There is one special problem when using alkyl-β-D-glucopyranosides or alkyl-β-D-maltosides as detergents: the commercially available detergents are often 'contaminated' with the α-anomers in varying, sometimes substantial, concentrations. The α-anomers are much less soluble, and appear to prevent crystallization. In the case of photosystem I from a thermophilic cyanobacterium, it has been reported that for a 2% α-anomer content in dodecyl-β-D-maltoside preparations no crystals can be obtained, with a 0.5–2% content the diffraction of the crystals is anisotropic with a reduction in resolution to 5–6 Å, whereas diffraction to better than 3.5 Å resolution in all directions is observed when the content of the α-anomer is below 0.1% (Fromme & Witt, 1998). The percentage of α-anomer can be determined using NMR spectroscopy or high-performance liquid chromatography. During ageing of detergent solutions, a conversion from the β-anomer to the α-anomer is expected. Therefore, ageing of detergent solutions has to be prevented.

The detergents that have been successfully used to crystallize membrane proteins can also be found in Table 4.2.1.1. The possibilities for developing new detergents for membrane-protein crystallization have not been exhausted. There is a need for new detergents, *e.g.* detergents with head groups with sizes between glucose and maltose are still missing!

It has been observed that crystals of the photosynthetic reaction centre from the purple bacterium *Rhodopseudomonas viridis* could be obtained when *N*,*N*-dimethyldodecylamine-*N*-oxide is used as detergent, but not when *N*,*N*-dimethyldecylamine-*N*-oxide is used. Even crystals formed with the dodecyl homologue lost their order when soaked in a buffer containing the decyl homologue. These observations were found to have an obvious explanation when the location of the detergent molecules bound to the protein was determined using neutron crystallography (Roth *et al.*, 1989); the detergent molecules surrounding neighbouring photosynthetic reaction centres in the crystal lattice are in contact. It is likely that attractive interactions between neighbouring protein-bound detergent micelles contribute to the stability of the crystal lattice. Particularly striking (see Table 4.2.3.1) is the dependence of the crystal quality on the alkyl-chain length in the case of the two-subunit cytochrome *c* oxidase from the soil bacterium *Paracoccus denitrificans.* Well ordered crystals were obtained with undecyl-β-D-maltoside, but not with the decyl and dodecyl homologues. Table 4.2.3.1 also lists the names of important vendors of detergents.

4.2.4. The 'small amphiphile concept'

From the arguments and observations presented above, it is evident that the size and shape of the detergent micelle are very important in membrane-protein crystallization. Detergent micelles can be made smaller (and their curvatures changed) when small amphiphilic molecules like heptane-1,2,3-triol are

Table 4.2.3.1

Summary of the results of attempts to crystallize the two-subunit cytochrome *c* oxidase from the soil bacterium *Paracoccus denitrificans* using different detergents (after Ostermeier *et al.*, 1997)

The resolutions of the crystals obtained are given in parentheses. Abbreviations: C_{12}: dodecyl; C_{11}: undecyl; C_{10}: decyl; C_9: nonyl; CYMAL-6: (cyclohexyl)hexyl-β-D-maltoside; CYMAL-5: (cyclohexyl)pentyl-β-D-maltoside. x in E_x is the number of oxyethylene units in the alkylpolyoxyethylene detergents. Suppliers: A: Anatrace (Maumee, OH); B: Biomol; C: Calbiochem; F: Fluka.

Detergent	Supplier	Crystals?
C_{12}-β-D-maltoside	B	Yes (8 Å)
C_{11}-β-D-maltoside	B	Yes (2.5 Å)
C_{10}-β-D-maltoside	B	No
CYMAL-6	A	Yes (2.6 Å)
CYMAL-5	A	No
Dodecylsucrose	C	No
Decylsucrose	C	No
C_9-β-D-glucoside	C	No
$C_{12}E_8$	F	Yes (> 8 Å)
$C_{12}E_6$	F	No
$C_{12}E_5$	F	No
$C_{10}E_6$	F	No
$C_{10}E_5$	F	No

added (Timmins *et al.*, 1991; Gast *et al.*, 1994). These compounds form mixed micelles with detergents. When 10% (*w*/*v*) heptane-1,2,3-triol is added to 1% solutions of *N*,*N*-dimethyldodecylamine-*N*-oxide in water, the number of detergent molecules per micelle decreases from 69 to 34, ~23 heptane-1,2,3-triol molecules are incorporated and the radius of the micelle is reduced from 20.9 to 16.9 Å (Timmins *et al.*, 1994). This so-called 'small amphiphile concept' has been used successfully to crystallize bacterial photosynthetic reaction centres (Michel, 1982, 1983; Buchanan *et al.*, 1993), bacterial light-harvesting complexes (Michel, 1991; Koepke *et al.*, 1996) and other membrane proteins (see Table 4.2.1.1). The light-harvesting complexes from the purple bacterium *Rhodospirillum molischianum* yield an astonishing number of different crystal forms, but only one diffracts to high resolution. A large amount of heptane-1,2,3-triol had to be added to obtain this crystal form. As a result, heptane-1,2,3-triol itself reached supersaturation during crystallization of the protein. Normally, the protein crystallizes first and heptane-1,2,3-triol later. When heptane-1,2,3-triol crystals form, the protein crystals crack and eventually redissolve. A likely explanation is that the concentration of solubilized heptane-1,2,3-triol is reduced when it forms crystals and the mixed detergent–heptane-1,2,3-triol micelles lose heptane-1,2,3-triol and take up detergent molecules from solution. The micelles expand and the crystals crack. This behaviour caused severe problems when searching for heavy-atom derivatives. Small amphiphiles work well with the *N*,*N*-dimethylalkylamine-*N*-oxides and alkylglucopyranosides as detergents, but not with alkylmaltosides. The most successful small amphiphiles are heptane-1,2,3-triol and benzamidine.

4.2.5. Membrane-protein crystallization with the help of antibody Fv fragments

The number of detergents that can be tried is limited for rather unstable membrane proteins. For instance, the four-subunit cytochrome *c* oxidase from *P. denitrificans* is sufficiently stable only with dodecyl-β-D-maltoside as detergent. When other detergents are employed, subunits III and IV and some lipids are removed from subunits I and II. These lipids might contribute to the binding of the protein subunits III and IV to the central subunits I and II. Therefore, the size of the detergent micelles can not be varied by using different detergents. In order to obtain crystals, the size of the extramembraneous part of this important enzyme was enlarged by the binding of Fv fragments of monoclonal antibodies. For this purpose, monoclonal antibodies against the four-subunit cytochrome *c* oxidase were generated using the classical hybridoma technique. Then hybridoma cell lines producing conformation-specific antibodies were selected (such antibodies react positively in enzyme-linked immunosorbent assays, but negatively in Western blot assays). The cDNA strands coding for the respective VL and VH genes were cloned and expressed in *Escherichia coli.* Binding of conformation-specific antibody fragments can be expected to lead to a more homogeneous protein preparation. The first Fv fragment worked, and well ordered crystals of the four-subunit and two-subunit cytochrome *c* oxidases were obtained (Ostermeier *et al.*, 1995, 1997). As an important additional advantage of this approach, an affinity tag can be fused to the recombinant antibody Fv fragment and used for efficient isolation. The affinity tag can then be used to purify the complex of antibody fragment and membrane protein rapidly and with a high yield (Kleymann *et al.*, 1995). Because Fv fragment cocrystallization also worked with the yeast cytochrome bc_1 complex, again with the first and only conformation-specific antibody Fv fragment tried (Hunte *et al.*, 2000), we are rather confident about this method for the future. The usefulness of single-chain Fv fragments, which can be obtained by the phage-display technique, has not been investigated, because the linker region joining the VL and VH chains must be expected to be flexible. Such flexibility would induce inhomogeneity and reduce the chance of obtaining crystals.

4.2.6. Membrane-protein crystallization using cubic bicontinuous lipidic phases

Landau & Rosenbusch (1996) introduced the use of bicontinuous cubic phases of lipids for membrane-protein crystallization. In such phases, the lipid forms a single, curved, continuous three-dimensional bilayer [see Lindblom & Rilfors (1989) for a review]. One can incorporate membrane proteins into such a bilayer, as demonstrated with octyl-β-D-glucopyranoside-solubilized monomeric bacteriorhodopsin. The three-dimensional bilayer network serves as a matrix for crystallization. The membrane protein can diffuse through the bilayer, but is also able to establish polar contacts in the third dimension. Landau & Rosenbusch demonstrated that bacteriorhodopsin forms small, well ordered three-dimensional crystals. The X-ray data indicate that the same two-dimensional crystals are present as formed by bacteriorhodopsin in its native environment (the purple membrane). These membranes are now stacked in the third dimension in a well ordered manner. Therefore, these crystals belong to type I.

The method has the conceptual problem that the growing three-dimensional crystal has to disrupt and displace the cubic lipidic phase. Nevertheless, it is hoped that this method can also be used for membrane proteins that do not have a strong tendency to form two-dimensional crystals spontaneously. In particular, this method appears to be the only chance for crystallizing those membrane proteins that are unstable in the absence of added lipids.

4.2.7. General recommendations

When trying to crystallize a membrane protein the first, and most important, task is to obtain a pure, stable and homogeneous preparation of the membrane protein under investigation.

Unfortunately, general methods for producing substantial amounts of membrane proteins using heterologous expression systems in a native membrane environment do not exist, nor is refolding of membrane proteins from inclusion bodies well established. Once a pure and homogeneous preparation of a membrane protein has been obtained, its stability in various detergents has to be investigated at different pH values. Frequently, a sharp optimum pH for stability is observed with detergents of shorter alkyl-chain length. In crystallization attempts, the usual parameters (nature of the precipitant, buffer, pH, addition of inhibitors and/or substrates, see Chapter 4.1) should be varied. The most important variable, however, is the detergent. Detergents can be exchanged most conveniently by binding the membrane protein to column materials (*e.g.* ion-exchange resins, hydroxyapatite, affinity matrices), washing with a buffer containing the new detergent at concentrations above the CMC under non-eluting conditions, and then eluting the bound membrane protein with a buffer containing the new detergent. The completeness of the detergent exchange should be checked. For this purpose, the use of ^{14}C- or ^{3}H-labelled detergents is recommended. Washing with about 20 column volumes is frequently required for a complete detergent exchange. It is more difficult to exchange a detergent with a low CMC (this means long alkyl chains) against a detergent with a high CMC (shorter alkyl chains). The detergent can often be exchanged during a final step in the purification. Less satisfying methods for detergent exchange are molecular-sieve chromatography, ultracentrifugation or repeated ultrafiltration and dilution. If the amount of membrane protein is limited, it is advisable to restrict the usual parameters to a set of, *e.g.*, 48 standard combinations with respect to precipitating agent, pH, buffer *etc.*, but to try *all* available detergents. If crystals of insufficient quality or size are obtained, trying the antibody Fv fragment approach is highly recommended. Alternatives are to use a different source (species) for the membrane protein under investigation, or to remove flexible parts of the membrane protein by proteolytic digestion. I am convinced that 50% of all membrane proteins will yield well ordered, three-dimensional crystals within five years, once the problem of obtaining a pure, stable and homogeneous preparation has been solved.

References

Allen, J. P., Feher, G., Yeates, T. O., Komiya, H. & Rees, D. C. (1987). *Structure of the reaction center from Rhodobacter sphaeroides R-26: the protein subunits. Proc. Natl Acad. Sci. USA*, **84**, 6162–6166.

Bordier, C. (1981). *Phase separation of integral membrane proteins in Triton X-114 solution. J. Biol. Chem.* **256**, 1604–1607.

Buchanan, S. K., Fritzsch, G., Ermler, U. & Michel, H. (1993). *New crystal form of the photosynthetic reaction centre from Rhodobacter sphaeroides of improved diffraction quality. J. Mol. Biol.* **230**, 1311–1314.

Buchanan, S. K., Smith, B. S., Venkatrami, L., Xia, D., Esser, L., Palnitkar, M., Chakraborty, R., van der Helm, D. & Deisenhofer, J. (1999). *Crystal structure of the outer membrane active transporter FepA from Escherichia coli. Nature Struct. Biol.* **6**, 56–63.

Chang, C. H., El-Kabbani, D., Tiede, D., Norris, J. & Schiffer, M. (1991). *Structure of the membrane-bound protein photosynthetic reaction center from Rhodobacter sphaeroides. Biochemistry*, **30**, 5352–5360.

Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. (1998). *Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel. Science*, **282**, 2220–2226.

Cowan, S. W., Garavito, R. M., Jansonius, J. N., Jenkins, J. A., Karlsson, R., König, N., Pai, E. F., Pauptit, R. A., Rizkallah, P. J., Rosenbusch, J. P., Rummel, G. & Schirmer, T. (1995). *The structure of OmpF porin in a tetragonal crystal form. Structure*, **3**, 1041–1050.

Cowan, S. W., Schirmer, T., Rummel, G., Steiert, M., Gosh, R., Pauptit, R. A., Jansonius, J. N. & Rosenbusch, J. P. (1992). *Crystal structures explain functional properties of two E. coli porins. Nature (London)*, **358**, 727–733.

Deisenhofer, J., Epp, O., Miki, K., Huber, R. & Michel, H. (1985). *Structure of the protein subunits in the photosynthetic reaction center of Rhodopseudomonas viridis at 3 Å. Nature (London)*, **318**, 618–642.

Deisenhofer, J., Epp, O., Sinning, I. & Michel, H. (1995). *Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from Rhodopseudomonas viridis. J. Mol. Biol.* **246**, 429–457.

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A. L., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). *The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. Science*, **280**, 69–77.

Ermler, U., Fritzsch, G., Buchanan, S. K. & Michel, H. (1994). *Structure of the photosynthetic reaction centre from Rhodobacter sphaeroides at 2.65 Å resolution: cofactors and protein–cofactor interactions. Structure*, **2**, 925–936.

Essen, L. O., Siegert, R., Lehmann, W. D. & Oesterhelt, D. (1998). *Lipid patches in membrane protein oligomers – crystal structure of the bacteriorhodpsin–lipid complex. Proc. Natl Acad. Sci. USA*, **95**, 11673–11678.

Ferguson, A. D., Hofmann, E., Coulton, J. W., Diederichs, K. & Welte, W. (1998). *Siderophore mediated iron transport: crystal structure of FhuA with bound lipopolysaccharide. Science*, **282**, 2215–2220.

Forst, D., Welte, W., Wacker, T. & Diederichs, K. (1998). *Structure of the sucrose-specific porin ScrY from Salmonella typhimurium and its complex with sucrose. Nature Struct. Biol.* **5**, 37–46.

Fromme, P. & Witt, H. T. (1998). *Improved isolation and crystallization of photosystem I for structural analysis. Biochim. Biophys. Acta*, **1365**, 175–184.

Gast, P., Hemelrijk, P. & Hoff, A. J. (1994). *Determination of the number of detergent molecules associated with the reaction center protein isolated from the photosynthetic bacterium Rhodopseudomonas viridis. Effects of the amphiphilic molecule, 1,2,3-heptanetriol. FEBS Lett.* **337**, 39–42.

Gerstein, M. (1998). *Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. Proteins*, **33**, 518–534.

Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. (1996). *Electron crystallographic refinement of the structure of bacteriorhodopsin. J. Mol. Biol.* **259**, 393–421.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). *Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. J. Mol. Biol.* **213**, 899–929.

Henderson, R. & Shotton, D. (1980). *Crystallization of purple membrane in three dimensions. J. Mol. Biol.* **139**, 99–109.

Hirsch, A., Breed, J., Saxena, K., Richter, O. M. H., Ludwig, B., Diederichs, K. & Welte, W. (1997). *The structure of porin from Paracoccus denitrificans at 3.1 Å resolution. FEBS Lett.* **404**, 208–210.

Hjelmeland, L. M. (1990). *Solubilization of native membrane proteins. Methods Enzymol.* **182**, 253–264.

Hunte, C., Lange, C., Koepke, J., Rossmanith, T. & Michel, H. (2000). *Structure at 2.3 Å resolution of the cytochrome bc_1 complex from the yeast Saccharomyces cerevisiae co-crystallized with an antibody Fv fragment. Structure*, **8**, 669–684.

Iwata, S., Lee, J. W., Okada, K., Lee, J. K., Iwata, M., Rasmussen, B., Link, T. A., Ramaswamy, S. & Jap, B. K. (1998). *Complete structure of the 11-subunit bovine mitochondrial cytochrome bc_1 complex. Science*, **281**, 64–71.

Iwata, S., Ostermeier, C., Ludwig, B. & Michel, H. (1995). *Structure at 2.8 Å resolution of cytochrome c oxidase from Paracoccus denitrificans. Nature (London)*, **376**, 660–669.

Kim, H., Xia, D., Yu, C. A., Xia, J. Z., Kachurin, A. M., Li, Z., Yu, L. & Deisenhofer, J. (1998). *Inhibitor binding changes domain mobility in the iron–sulfur protein of the mitochondrial bc_1 complex from bovine heart. Proc. Natl Acad. Sci. USA*, **95**, 8026–8033.

Kimura, Y., Vassylyev, D. G., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T. & Fujiyoshi, Y. (1997). *Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. Nature (London)*, **389**, 206–211.

Kleymann, G., Ostermeier, C., Ludwig, B., Skerra, A. & Michel, H. (1995). *Engineered F_v fragments as a tool for the one-step purification* *of integral multisubunit membrane protein complexes. Biotechnology*, **13**, 155–160.

Koepke, J., Hu, X., Muenke, C., Schulten, K. & Michel, H. (1996). *The crystal structure of the light-harvesting complex II (B800–850) from Rhodospirilum molischianum. Structure*, **4**, 581–597.

Kreusch, A., Neubüser, A., Schiltz, E., Weckesser, J. & Schulz, G. E. (1994). *Structure of the membrane channel porin from Rhodopseudomonas blastica at 2.0 Å resolution. Protein Sci.* **3**, 58–63.

Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. (1994). *Atomic model of plant light-harvesting complex by electron crystallography. Nature (London)*, **367**, 614–621.

Kurumbail, R. G., Stevens, A. M., Gierse, J. K., McDonald, J. J., Stegeman, R. A., Pak, J. Y., Gildehaus, D., Miyashiro, J. M., Penning, T. D., Seibert, K., Isakson, P. C. & Stallings, W. C. (1996). *Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. Nature (London)*, **384**, 644–648.

Lancaster, C. R. D. & Michel, H. (1997). *The coupling of light-induced electron transfer and proton uptake as derived from crystal structures of reaction centres from Rhodopseudomonas viridis modified at the binding site of the secondary quinone, Q_B. Structure*, **5**, 1339–1359.

Lancaster, C. R. D. & Michel, H. (1999). *Refined crystal structures of reaction centres from Rhodopseudomonas viridis in complexes with the herbicide atrazine and two chiral atrazine derivatives also lead to a new model of the bound carotenoid. J. Mol. Biol.* **286**, 883–898.

Landau, E. M. & Rosenbusch, J. P. (1996). *Lipidic cubic phases: a novel concept for the crystallization of membrane proteins. Proc. Natl Acad. Sci. USA*, **93**, 14532–14535.

Lindblom, G. & Rilfors, L. (1989). *Cubic phases and isotropic structures formed by membrane lipids – possible biological relevance. Biochim. Biophys. Acta*, **988**, 221–256.

Locher, K. P., Rees, B., Koebnik, R., Mitschler, A., Moulinier, L., Rosenbusch, J. P. & Moras, D. (1998). *Transmembrane signaling across the ligand-gated FhuA receptor: crystal structures of free and ferrichrome-bound states reveal allosteric changes. Cell*, **98**, 771–778.

Luecke, H., Richter, H. T. & Lamyi, J. K. (1998). *Proton transfer pathways in bacteriorhodopsin at 2.3 Å resolution. Science*, **280**, 1934–1937.

Luong, C., Miller, A., Barnett, J., Chow, J., Ramesha, C. & Browner, M. F. (1996). *Flexibility of the NSAID binding site in the structure of human cyclooxygenase-2. Nature Struct. Biol.* **3**, 927–933.

McDermott, G., Prince, S. M., Freer, A. A., Hawthornthwaite-Lawless, A. M., Papiz, M. Z., Cogdell, R. J. & Isaacs, N. W. (1995). *Crystal-structure of an integral membrane light-harvesting complex from photosynthetic bacteria. Nature (London)*, **374**, 517–521.

Meyer, J. E. W., Hofnung, M. & Schulz, G. E. (1997). *Structure of maltoporin from Salmonella typhimurium ligated with a nitrophenyl-maltotrioside. J. Mol. Biol.* **266**, 761–775.

Michel, H. (1982). *Three-dimensional crystals of a membrane protein complex. The photosynthetic reaction centre from Rhodopseudomonas viridis. J. Mol. Biol.* **158**, 567–572.

Michel, H. (1983). *Crystallization of membrane proteins. Trends Biochem. Sci.* **8**, 56–59.

Michel, H. (1991). Editor. *Crystallization of Membrane Proteins.* Boca Raton, Florida: CRC Press.

Midura, R. J. & Yanagishita, M. (1995). *Chaotropic solvents increase the critical micellar concentrations of detergents. Anal. Biochem.* **228**, 318–322.

Neugebauer, J. M. (1990). *Detergents: an overview. Methods Enzymol.* **182**, 239–253.

Ostermeier, C., Harrenga, A., Ermler, U. & Michel, H. (1997). *Structure at 2.7 Å resolution of the Paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody F_v fragment. Proc. Natl Acad. Sci. USA*, **94**, 10547–10553.

Ostermeier, C., Iwata, S., Ludwig, B. & Michel, H. (1995). *F_v fragment-mediated crystallization of the membrane protein bacterial cytochrome c oxidase. Nature Struct. Biol.* **2**, 842–846.

Pautsch, A. & Schulz, G. E. (1998). *Structure of the outer membrane protein: a transmembrane domain. Nature Struct. Biol.* **5**, 1013–1017.

Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. (1997). *X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipidic cubic phases. Science*, **277**, 1676–1681.

Picot, D., Loll, P. J. & Garavito, M. (1994). *The X-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. Nature (London)*, **367**, 243–249.

Roth, M., Lewit-Bentley, A., Michel, H., Deisenhofer, J., Huber, R. & Oesterhelt, D. (1989). *Detergent structure in crystals of a bacterial photosynthetic reaction center. Nature (London)*, **340**, 659–662.

Schirmer, T., Keller, T. A., Wang, Y. F. & Rosenbusch, J. P. (1995). *Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. Science*, **267**, 512–514.

Song, L., Hobaugh, M. R., Shustak, C., Cheley, S., Bayley, H. & Gouaux, J. E. (1996). *Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. Science*, **274**, 1859–1866.

Stowell, M. H. B., McPhillips, T. M., Rees, D. C., Soltis, S. M., Abresch, E. & Feher, G. (1997). *Light-induced structural changes in photosynthetic reaction center: implications for mechanism of electron–proton transfer. Science*, **276**, 812–816.

Timmins, P. A., Hauk, J., Wacker, T. & Welte, W. (1991). *The influence of heptane-1,2,3-triol on the size and shape of LDAO micelles. Implications for the crystallization of membrane proteins. FEBS Lett.* **280**, 115–120.

Timmins, P. A., Pebay-Peyroula, E. & Welte, W. (1994). *Detergent organisation in solutions and in crystals of membrane proteins. Biophys. Chem.* **53**, 27–36.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1995). *Structures of metal sites of oxidized bovine heart cytochrome c oxidase at 2.8 Å. Science*, **269**, 1069–1074.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). *The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Science*, **272**, 1136–1144.

Weiss, M. S., Abele, U., Weckesser, J., Welte, W., Schiltz, E. & Schulz, G. E. (1991). *Molecular architecture and electrostatic properties of a bacterial porin. Science*, **254**, 1627–1630.

Wendt, K. U., Poralla, K. & Schulz, G. E. (1997). *Structure and function of a cyclase. Science*, **277**, 1811–1815.

Xia, D., Yu, C. A., Kim, H., Xia, J. Z., Kachurin, A. M., Zhang, L., Yu, L. & Deisenhofer, J. (1997). *Crystal structure of the cytochrome bc_1 complex from bovine heart mitochondria. Science*, **277**, 60–66.

Yoshikawa, S., Shinzawa-Itoh, K., Nakashima, R., Yaono, R., Yamashita, E., Inoue, N., Yao, M., Fei, M. J., Libeu, C. P., Mizushima, T., Yamaguchi, H., Tomizaki, T. & Tsukihara, T. (1998). *Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science*, **280**, 1723–1729.

Zhang, Z. L., Huang, L. S., Shulmeister, V. M., Chi, Y. I., Kim, K. K., Hung, L. W., Crofts, A. R., Berry, E. A. & Kim, S. H. (1998). *Electron transfer by domain movement in cytochrome bc_1. Nature (London)*, **392**, 677–684.

Zulauf, H. (1991). *Detergent phenomena in membrane protein crystallization.* In *Crystallization of Membrane Proteins*, edited by H. Michel, ch. 2, pp. 53–72. Boca Raton, Florida: CRC Press.

International Tables for Crystallography (2012). Vol. F, Chapter 4.3, pp. 129–139.

Chapter 4.3. Application of protein engineering to enhance crystallizability and improve crystal properties

Z. S. Derewenda

4.3.1. Introduction

Following the dawn of recombinant technology brought about by the groundbreaking overexpression of synthetic genes coding for insulin and somatostatin in *Escherichia coli* (Goeddel *et al.*, 1979; Itakura *et al.*, 1977) and the subsequent discovery of the polymerase chain reaction (PCR; Mullis *et al.*, 1986; Saiki *et al.*, 1985, 1988), macromolecular crystallography was freed of its long-standing dependence on purified native protein samples for crystallization. Heterologous expression made it possible to generate samples of proteins and complexes that are found in only small or trace amounts in living cells and to engineer large and unstable proteins so that isolated domains or modified forms can be made available for crystallization. At the same time, the effort required for protein purification was dramatically reduced by the use of fusion proteins and affinity tags (Brewer *et al.*, 1991; Sassenfeld, 1990; Malhotra, 2009). As a consequence, the overwhelming majority of samples used today for crystallization are recombinantly derived proteins. However, even though material for crystallization is more easily available, the preparation of single, well diffracting crystals of the target macromolecule is still a time-consuming challenge.

Historically, two complementary approaches to protein crystallization were developed in parallel. Firstly, natural variations in the amino-acid sequences of homologues from different species were exploited to identify a target with suitable crystallization properties during the purification procedure (Kendrew *et al.*, 1954; Campbell *et al.*, 1972). The second approach was to extensively screen a specific target protein against a range of diverse precipitating agents, buffers and additives until the right conditions for crystallization were identified (Carter & Carter, 1979; Jancarik & Kim, 1991). These strategies remain the pillars of contemporary macromolecular crystallization. However, as the palette of molecular biology techniques expanded to include site-directed mutagenesis, ligation-independent cloning and other tools, it became possible to modify proteins with relative ease with the specific purpose of enhancing their propensity to crystallize or improving the diffraction quality of the resulting crystals. The early proof-of-principle of these capabilities was the crystallization of an engineered variant of human H-ferritin in which a single-site mutation, K86Q, was introduced to duplicate a crystal contact mediated by Cd^{2+} ions in the crystal structure of the homologous rat L-ferritin (Lawson *et al.*, 1991).

In this review, current progress in the methodologies of protein engineering used to enhance the crystallizability of targets that are recalcitrant to crystallization in their wild-type form is discussed. This burgeoning field is very broad and includes both general strategies that apply to a range of targets and many diverse approaches that only apply to specific proteins or protein families. Thus, owing to space limitations, the focus is on those techniques that have either been demonstrated to be of general utility or are at a point in their development to clearly have the potential to become widely used in the future. Understandably, only representative examples are provided.

4.3.2. Microscopic aspects of protein crystallization

Proteins are inherently dynamic entities, a property that greatly hinders their crystallizability. Not surprisingly, it has been estimated that even for the stable and relatively small single-domain prokaryotic proteins fewer than one in four will yield X-ray-quality crystals when using a routine screening process (Canaves *et al.*, 2004; Price *et al.*, 2009). In order to rationally modify proteins to enhance their crystallizability, it is first necessary to understand the physical properties that make most proteins resistant to crystallization.

Protein crystals are nucleated *ab initio* at supersaturation levels in the 200–1000% range (McPherson, 1982). Nucleation is believed to proceed *via* a two-step process: clusters of solute molecules form first and upon reaching critical size reorganize into three-dimensionally ordered nuclei (Georgalis *et al.*, 1997; Vekilov, 2004; Erdemir *et al.*, 2009). Subsequent transfer of protein molecules from solution onto the growing crystal surface is driven by relatively small negative changes in Gibbs free energy (ΔG°), from approximately -10 to $-100\,\mathrm{kJ\,mol^{-1}}$, at ambient temperature (Vekilov, 2003). Interestingly, enthalpy changes are generally negligible during crystallization (Yau *et al.*, 2000; Petsev *et al.*, 2001; Gliko *et al.*, 2005), so that entropic phenomena dominate (Vekilov *et al.*, 2002; Vekilov, 2003; Derewenda & Vekilov, 2006). The microscopic effects underlying the entropy changes, both favourable and unfavourable, involve the protein itself as well as the solvent. Protein packing, which results in an ordered three-dimensional lattice and loss of translational and rotational degrees of freedom, is unfavourable and produces an energy barrier in the 30–100 $\mathrm{kJ\,mol^{-1}}$ range at room temperature (Finkelstein & Janin, 1989; Tidor & Karplus, 1994). Similarly, incorporation into the growing crystal and ordering of any intrinsically unstructured elements, such as flexible termini or loops and side chains, at the point of crystal contacts further increases the entropic cost. However, the release of ordered solvent molecules from the surfaces involved in crystal contacts, which is estimated to be in the 25–150 $\mathrm{kJ\,mol^{-1}}$ range, may sufficiently compensate for these entropy losses and ultimately provide the driving force for crystal growth (Vekilov *et al.*, 2002; Vekilov, 2003).

Based on these considerations, it is evident that a protein must satisfy certain criteria in order to crystallize. Firstly, it must have a molecular surface that confers adequate solubility under initial conditions to reach the necessary supersaturation level for nucleation. Furthermore, it should have few, if any, intrinsically unstructured fragments such as extended N- or C-termini or long and solvent-exposed loops which may impede crystallization. Finally, the protein should have distinct ‘sticky’ patches on the surface with a structured layer of solvent molecules, allowing the ordering of nascent nuclei by mediating thermodynamically viable specific crystal contacts.

The notion that protein crystallization involves specific and anisotropic intermolecular interactions, as opposed to random contacts, is relatively new. Early analyses of intermolecular contacts in protein crystals concluded that crystallization is a

Copyright © 2012 International Union of Crystallography

stochastic process generated by mostly random contacts (Janin & Rodier, 1995; Janin, 1997; Carugo & Argos, 1997). However, more recent stringent statistical analyses using a larger database strongly suggested that crystal contacts are generated by anisotropic interactions that favour small hydrophobic residues and disfavour large polar side chains with high conformational entropy (Cieślik & Derewenda, 2009). This view is also supported by a large-scale comparison of the amino-acid sequences of crystallizable and noncrystallizable proteins, which established that crystallization propensity is negatively correlated with the prevalence of residues with high side-chain entropy (Price *et al.*, 2009). Finally, molecular-dynamics simulations of the intermolecular interactions of lysozyme in solution show that they are anisotropic and that their magnitude and nature depend on the physical chemistry of the participating interfaces, suggesting that the nucleation phenomenon is initiated in a nonstochastic fashion (Pellicane *et al.*, 2008).

Understanding the physical principles that govern crystallization at the microscopic level provides the singular underpinning to rationally engineer target proteins to enhance their crystallizability either by improving their solution properties or by increasing their propensity to engage in weak but specific interactions that organize the transformation of nascent clusters into nuclei and drive subsequent crystal growth.

4.3.3. Engineering proteins with enhanced solubility

The solubility of a protein is the primary essential prerequisite for its crystallization. It should be noted that the expression 'low solubility' is often used indiscriminately to describe quite different phenomena, including a propensity to aggregate and precipitate upon overexpression owing to misfolding, amyloid formation and finally genuine low *in vitro* solubility, *i.e.* low protein concentration in equilibrium with the solid phase, of otherwise fully folded and stable proteins (Trevino *et al.*, 2008). Here, the strategies and methods that specifically address the latter case are discussed, *i.e.* precipitation at low concentrations of properly folded proteins.

It has been well established that even single-site mutations of surface residues can dramatically affect the solubility of a protein and its crystallizability (McElroy *et al.*, 1992). Consequently, the intuitively obvious approach is to mutate solvent-exposed hydrophobic amino acids to hydrophilic residues. In this way, the low solubility of the catalytic domain of HIV-1 integrase was addressed by screening 29 mutants in which hydrophobic residues were systematically mutated to hydrophilic amino acids; of the variants tested, the single-site mutant F185K showed a dramatically improved solubility and ultimately yielded X-ray-quality crystals (Dyda *et al.*, 1994; Jenkins *et al.*, 1995). In the case of leptin, the product of the *obese* gene, the solubility-enhancing W100E mutation proved to be critical for crystallization of the protein (Zhang *et al.*, 1997). Recently, a screen of several variants of human apolipoprotein D identified a triple mutant (W99H, I118S, L120S) which was much more soluble than the wild-type protein and which was ultimately used to obtain well diffracting crystals (Nasreen *et al.*, 2006; Eichinger *et al.*, 2007).

While engineering enhanced solubility using site-directed mutagenesis is potentially a powerful approach, in the absence of structural information it is a challenge to predict which hydrophobic residues are solvent-exposed and might therefore constitute useful targets for mutagenesis. Moreover, even if structural information is available for a homologue or the target itself, it may not be clear what type of mutation actually works best, forcing the investigator to rely on extensive screening. This uncertainty arises from the fact that hydrophobicity scales for individual amino acids cannot be used directly to evaluate the increase or decrease of protein solubility as a consequence of a specific mutation. Furthermore, there have been few rigorous studies of the effects of specific mutations on protein solubility. A notable example is a study on ribonuclease SA in which the solvent-exposed Thr76 was replaced by 19 other amino acids and the solubility of all of the variants was carefully evaluated (Trevino *et al.*, 2007). Those variants that contained Asp, Arg, Glu and Ser were the most soluble. Unexpectedly, even though a lysine might be expected to confer higher solubility than a serine or alanine, the T76S mutation actually led to a significantly higher solubility than T76K, while the T76A variant was only marginally less soluble than T76K (Trevino *et al.*, 2007). The authors of the study concluded that mutating Asn and Gln to their respective acids may constitute the most robust strategy of enhancing solubility. Interestingly, one of the first examples of rational enhancement of solubility, *i.e.* the study of trimethoprim-resistant type S1 hydrofolate reductase (Dale *et al.*, 1994), used this very strategy: the amide-containing side chains were systematically substituted with carboxylic amino acids and one specific variant, a double mutant N48E, N130D, was found to exhibit markedly increased solubility and ultimately yielded crystals that were suitable for crystallographic analysis.

Somewhat ironically, large charged residues such as glutamate that confer higher solubility on the target protein may at the same time impede crystallization because they increase the total surface side-chain entropy, making the surface recalcitrant to engaging in crystal contact-mediating interactions. Thus, variants engineered for increased solubility may simultaneously show a decreased propensity to crystallize.

Some of the above uncertainties can be overcome with an alternative approach of directed evolution and phenotypic selection methods, in which soluble mutants are directly selected from vast protein libraries (Farinas *et al.*, 2001; Farinas, 2006; Pédelacq *et al.*, 2002; Waldo, 2003; Cabantous *et al.*, 2005). Several different variations of this method have been reported (Waldo, 2003). For example, the target protein may be fused to the N-terminus of a reporter protein such as the green fluorescent protein (GFP; Waldo *et al.*, 1999) or direct detection methods can be employed to identify soluble variants (Peabody & Al-Bitar, 2001). While elegant and potentially very effective, directed evolution has not yet been widely adopted for the generation of crystallizable proteins.

Solubility problems are not always caused by excessively exposed hydrophobic surfaces. In some cases, the root of the problem is aggregation caused by exposed free cysteines. Reduced cysteines can be identified by alkylation with *N*-ethylmaleimide or iodoacetamide under anaerobic conditions and subsequent electrospray mass spectrometry (Niessing *et al.*, 2004). Several examples illustrate how this approach is helpful in generating samples that are suitable for crystallization. In mitogen-activated protein (MAP) kinase p38α, a single-site mutation (C162S) prevented aggregation and yielded a crystallizable variant (Patel *et al.*, 2004). Similarly, a double mutant (C95K, C142S) of foot-and-mouth disease virus 3C protease showed none of the aggregation problems that plagued the wild-type protein and was subsequently crystallized (Birtley & Curry, 2005). It is noteworthy that in this case an alternative strategy involving mutations of the exposed hydrophobic residues Met81, Leu82 and Val140 did not eliminate aggregation (Birtley & Curry, 2005). In a number of cases aggregation problems were traced to

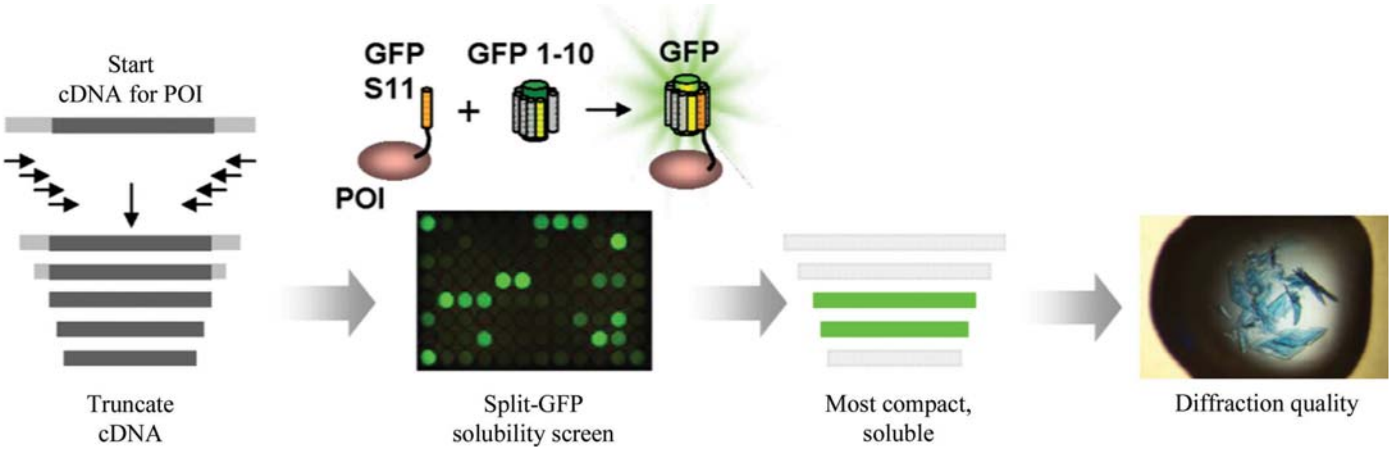


Figure 4.3.4.1
A domain-trapping strategy to engineer soluble variants of a protein of interest (POI) for crystallization using the split-GFP complementation methodology. (Figure courtesy of Dr Geoff Waldo, LANL.)

multiple free cysteines. In She2p, an RNA-binding protein, four cysteines (Cys14, Cys68, Cys106 and Cys180) were mutated to serines in order to overcome oxidation and aggregation (Niessing *et al.*, 2004). In an extreme case, that of human maspin, which is a serpin with antitumour activities, all unpaired cysteines were mutated (C20S, C34A, C183S, C205S, C214S, C297S, C373S) in an effort to obtain a soluble crystallizable variant (Al-Ayyoubi *et al.*, 2004).

4.3.4. Optimization of target constructs

The N- and C-termini of proteins are often flexible and unstructured (Thornton & Sibanda, 1983), creating a potential entropic impediment to crystallization. Initially, the preferred way to circumvent this problem was to use limited proteolysis to trim off the ends, leaving the stable core of the target protein. This strategy remains useful, particularly in its *in situ* version, in which trace amounts of proteases are added directly to crystallization screens (Dong *et al.*, 2007; Wernimont & Edwards, 2009). However, on the downside it introduces the possibility of heterogeneity in the sample owing to incomplete proteolysis. An alternative route is to first identify the smallest functional fragment of the target protein and to then design and overexpress an appropriately modified gene. A number of options are possible. The simplest is the direct prediction of intrinsically disordered regions from the amino-acid sequence alone (Obradovic *et al.*, 2003; He *et al.*, 2009). The functional core units can also be identified experimentally by mass spectrometry following limited proteolysis (Cohen *et al.*, 1995). Alternatively, deuterium–hydrogen exchange coupled to mass spectrometry (DXMS) may be used to identify fast-exchanging amides that map to unstructured fragments (Hamuro *et al.*, 2003; Pantazatos *et al.*, 2004; Sharma *et al.*, 2009).

Importantly, the choice of optimal N- and C-termini may also critically influence the solubility of the target protein. For example, in the case of MAPKAP kinase 2, 16 truncation variants were assayed, all of which contained the catalytic domain, and were shown to have dramatically differing solubilities and propensities for crystallization (Malawski *et al.*, 2006). Similarly, a series of truncations were screened in order to identify a soluble and crystallizable variant of a three-domain fragment of the Vav1 guanine nucleotide-exchange factor (Brooun *et al.*, 2007). In both these cases only a limited number of rationally designed constructs were screened. However, to increase the prospects of success it is also possible to utilize much larger libraries of variants and screen them *in vivo* using the high-throughput split-GFP complementation assay (Fig. 4.3.4.1; Cabantous & Waldo, 2006).

Another troublesome problem associated with flexible termini is their occasional propensity to form multiple intermolecular contacts, leading to crystal forms that contain multiple copies of the target protein in the asymmetric unit. This has been observed, for example, for *Plasmodium falciparum* peptide deformylase, in which removal of three residues from the N-terminus reduced the number of subunits in the asymmetric unit from ten to two (Robien *et al.*, 2004).

In addition to the disordered N- and C-termini, target proteins may contain internal unstructured regions such as subdomains or loops which can be removed or shortened to reduce conformational heterogeneity. For example, the construct used in the successful crystallization of the HIV gp120 envelope glycoprotein had two flexible loops which were replaced with Gly-Ala-Gly linkages to obtain a crystallizable variant (Kwong *et al.*, 1998, 1999). In the case of 8*R*-lipoxygenase the replacement of a flexible Ca^{2+}-dependent membrane-insertion loop consisting of five amino acids by a Gly-Ser dipeptide resulted in crystals that diffracted to a resolution 1 Å higher than the wild-type protein (Neau *et al.*, 2007). An interesting variation of this approach was introduced for the preparation of crystals of the β-subunit of the signal recognition particle receptor. A 26-residue flexible loop was removed, but instead of replacing it with a shorter sequence the authors connected the native N- and C-termini of the protein using a heptapeptide GGGSGGG, thus creating a circular permutation of the polypeptide chain (Schwartz *et al.*, 2004).

Given that the majority of eukaryotic proteins contain at least one stretch of 40 or more disordered residues (Vucetic *et al.*, 2003), optimization of crystallization targets by removal of these sequences is likely to become a routine strategy.

4.3.5. The use of fusion proteins for crystallization

Tags are routinely used in heterologous protein expression in order to enhance folding and solubility and to facilitate purification (Uhlen *et al.*, 1992; Malhotra, 2009). They are either short oligopeptides, such as a hexahistidine, with unique affinity properties or well expressed and highly soluble proteins, such as GST (glutathione *S*-transferase), MBP (maltose-binding protein) or thioredoxin. The tags are inserted into the expression vectors

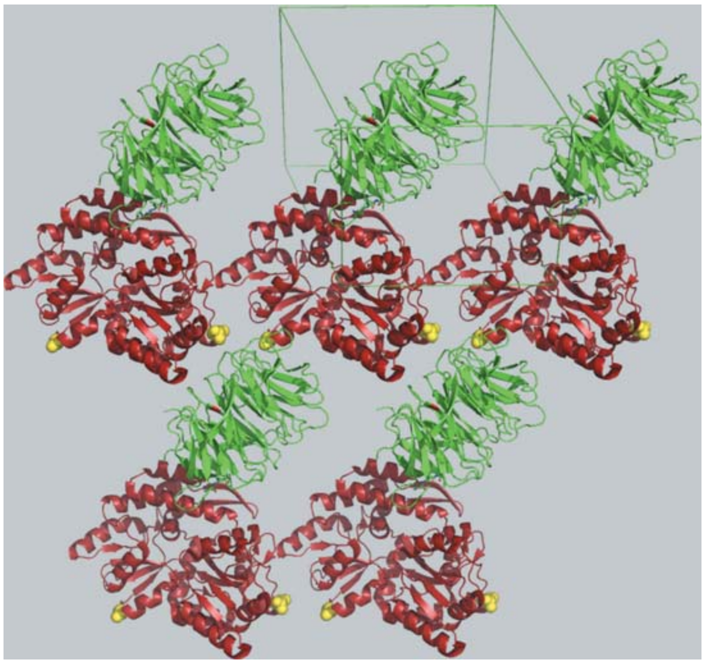


Figure 4.3.5.1
An example of the use of a fused carrier protein in crystallization: the crystal structure of the RACK1 protein (green) crystallized in fusion with an engineered variant of the maltose-binding protein (MBP; red); the major crystal contacts are mediated by MBP (PDB code 3dm0; Ullah *et al.*, 2008). The yellow spheres show alanines introduced by site-directed mutagenesis (see text for further details). Figs. 4.3.5.1 and 4.3.9.1 were generated using *PyMOL* (http://www.pymol.org).

downstream or upstream of the target protein and are often separated from it by a protease-sensitive linker sequence. They are cleaved proteolytically following expression and partial purification of the fusion protein and removed, leaving the isolated target ready for crystallization. However, in some cases the target protein may not be adequately soluble after cleavage or may resist crystallization. One of the possible solutions is to use the intact fusion protein in the crystallization screens in the hope that the carrier protein will both confer solubility on the construct and mediate crystal contacts. Not surprisingly, the canonical carrier proteins, all of which crystallize fairly easily on their own, constitute the obvious first choice. Using this strategy, the DNA-binding domain of DNA replication-related element-binding factor (DREF) was crystallized in fusion with *Escherichia coli* GST (Kuge *et al.*, 1997) and the U2AF homology motif (UHM) domain of splicing factor Puf60 was crystallized as a fusion with thioredoxin (Corsini *et al.*, 2008). A key problem limiting the utility of this technique is the inherent flexibility of a two-domain fusion protein, which is detrimental to its crystallizability. A possible solution to this problem is shortening the linker between the two proteins until a relatively rigid construct is identified (Smyth *et al.*, 2003). This approach was successfully pioneered for maltose-binding protein (MBP), which was used as a fusion chaperone to crystallize the human T-cell leukemia virus type 1 gp21 ectodomain fragment (Center *et al.*, 1998). The same strategy was employed in the crystallization of the ZP-N domain of ZP3 (Monne *et al.*, 2008), the islet amyloid polypeptide (IAPP; Wiltzius *et al.*, 2009) and the MATα1 homeodomain (Ke & Wolberger, 2003). Recently, a genetically modified version of MBP (see below) was used as an N-terminal fusion chaperone to crystallize the signal transduction regulator RACK1 from *Arabidopsis thaliana* (Ullah *et al.*, 2008; Fig. 4.3.5.1). Thus, MBP remains the most successful fusion chaperone for protein crystallization, even though the absolute number of proteins crystallized in this way is still limited.

In addition to the canonical fusion chaperones, which were originally designed as affinity tags, other carrier proteins can be used to assist crystallization. For example, a module made up of two sterile α motif (SAM) domains has been engineered to polymerize in response to a pH drop and was shown to drive the crystallization of 11 target proteins in a pilot study (Nauli *et al.*, 2007). In another example, barnase, a secreted ribonuclease from *Bacillus amyloliquefaciens*, was recently used as a carrier protein for crystallization of the disulfide-rich protein McoEeTI (Niemann *et al.*, 2006).

An alternative to N- or C-terminal fusions is an insertion fusion, in which a carrier protein is inserted into a loop in the sequence of a poorly soluble target. To date, this approach has exclusively been used in membrane-protein crystallization and was initially pioneered for the *E. coli* lactose permease, in which cytochrome b_{562}, flavodoxin and T4 lysozyme were tested as carrier proteins inserted into one of the loops (Privé *et al.*, 1994; Engel *et al.*, 2002). In this specific case none of these variants actually yielded useful crystals and the structure of lactose permease was eventually solved using crystals obtained using a variant containing the C154G mutation which stabilized a single conformation in complex with a lactose analogue (Abramson *et al.*, 2003). In contrast, a similar insertion fusion with T4 lysozyme replacing the third intracellular loop of the β2-adrenergic receptor was highly successful and yielded good-quality crystals that allowed structure determination at 2.4 Å resolution (Cherezov *et al.*, 2007; Rosenbaum *et al.*, 2007). This spectacular result attests to the potential of insertion-fusion proteins, but the method is not trivial as the constructs must be carefully evaluated for both structural and functional consequences of the insertion and a number of variants may have to be screened before a suitable one is identified.

4.3.6. Noncovalent crystallization chaperones

Noncovalent crystallization chaperones, *i.e.* engineered binding proteins that produce noncovalent complexes with target macromolecules, constitute an exciting alternative to fusion carrier proteins. Complexes with such chaperones often exhibit enhanced solubility and/or crystallizability in comparison to the isolated targets. The Fab and Fv fragments of antibodies are most commonly used for this purpose (Kovari *et al.*, 1995; Hunte & Michel, 2002; Prongay *et al.*, 1990; Ostermeier *et al.*, 1995; Jiang *et al.*, 2003; Dutzler *et al.*, 2003; Lee *et al.*, 2005). In its canonical version, the technique requires animal immunization with subsequent purification of hybridoma-derived antibodies and their proteolytic digestion to obtain pure homogeneous Fab fragments (Karpusas *et al.*, 2001; Kovari *et al.*, 1995). Alternatively, the Fab fragment can be directly sequenced and a synthetic gene can be used for *E. coli* expression, although this is not trivial owing to the presence of disulfides and two separate polypeptide chains in an Fab molecule. To overcome this bottleneck, a more efficient method of recombinant production of antibody fragments using mammalian HEK 293T has recently been proposed (Nettleship *et al.*, 2008). Another possibility is the use of so-called nanobodies, *i.e.* single-chain fragments derived from camelid antibodies (Koide, Tereshko *et al.*, 2007; Lam *et al.*, 2009; Korotkov *et al.*, 2009). However, this strategy requires immunization of camels or llamas, which is not technically easy.

Regardless of the specific strategy, the use of hybridoma technology and animal immunization is always time-consuming

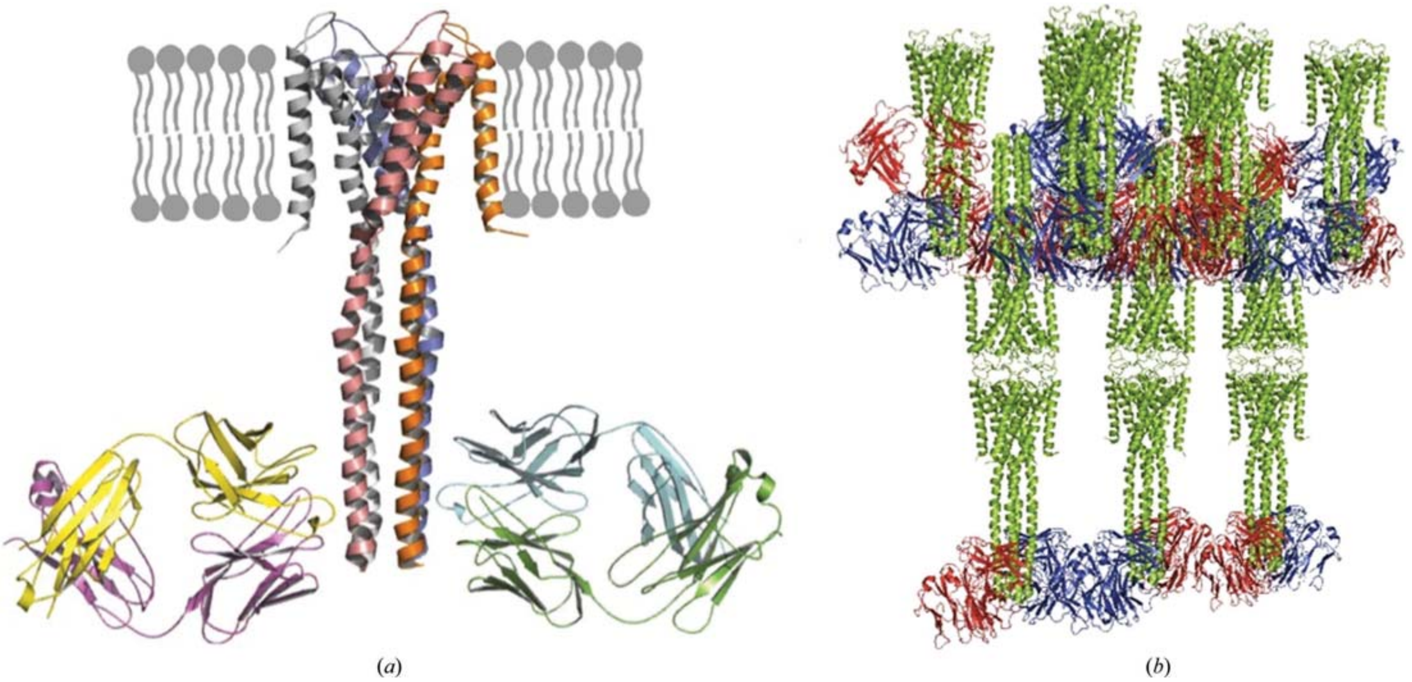


Figure 4.3.6.1
Phage-display-generated Fab fragments as crystallization chaperones: the structure of the KcsA channel in the closed conformation in complex with a synthetic Fab (PDB code 3eff; Uysal *et al.*, 2009). (*a*) A diagram showing how the Fab binds to the cytosolic portion in reference to the transmembrane domain. (*b*) The crystal structure of the complex showing how the synthetic Fab molecules mediate the major crystal contacts. (Figure courtesy of Dr Anthony Kossiakoff, University of Chicago.)

and expensive. In principle, a more efficient approach is to carry out *in vitro* selection of Fab fragments using phage display (Lee *et al.*, 2004) or ribosome display (Lipovsek & Pluckthun, 2004). However, since a typical antibody–antigen interface involves ~30 amino acids, the total number of possible sequences of a given template Fab significantly exceeds the available combinatorial libraries. Consequently, traditional phage-display libraries greatly diminish diversity at the mutated sites, which explains why synthetic antibodies were initially weaker binders than natural ones (Hawkins *et al.*, 1992; Koide, 2009). This problem was successfully overcome using a different type of phage-display library based on a 'reduced genetic code' and comprised of only a few amino acids, *e.g.* four, which produces high-affinity binders based on a single Fab scaffold (Fellouse *et al.*, 2004; Lee *et al.*, 2004). In contrast to natural antibodies, such synthetic Fab fragments can be generated against unique conformations, complexes or weak antigens such as RNA. Among recent examples are the crystallization and structure determination of the closed form of the full-length KcsA potassium channel with its cognate synthetic Fab (Uysal *et al.*, 2009; Fig. 4.3.6.1) and the crystallographic study of the ΔC209 P4-P6 domain of the *Tetrahymena* group I intron, a structured RNA molecule (Ye *et al.*, 2008).

The *in vitro* display methods also allow the engineering of non-antibody scaffolds as alternative protein binders and crystallization chaperones (Koide, 2009). For example, a fibronectin type III domain (FN3) scaffold was successfully used to generate binders with a reduced genetic code phage-display library (Koide, Gilbreth *et al.*, 2007; Gilbreth *et al.*, 2008). A similar approach was used for DARPins, *i.e.* designed ankyrin-repeat proteins (Sennhauser & Grütter, 2008), based on ribosome-display selection (Lipovsek & Pluckthun, 2004; Sennhauser & Grütter, 2008). Several new protein structures have been solved as complexes with DARPin chaperones, including polo-like kinase 1 (Bandeiras *et al.*, 2008), the trimeric integral membrane multidrug transporter AcrB (Sennhauser *et al.*, 2007) and the receptor-binding protein (RBP, the BppL trimer) of the baseplate complex of the lactococcal phage TP901-1 (Veesler *et al.*, 2009).

4.3.7. Removal of post-translational modifications

A number of proteins undergo post-translational modifications which can adversely affect crystallization. By far the most ubiquitous is N- and O-glycosylation, primarily of membrane-associated, secreted and lysosomal proteins. In a number of cases successful crystallization of glycoproteins purified from natural sources has been reported and carbohydrate groups have often been found to be ordered and occasionally sequestered between the protein molecules, thus even contributing in a positive way to crystallization (Mark *et al.*, 2003; Aleshin *et al.*, 1994). In general terms, however, the flexible and heterogeneous carbohydrate moieties, particularly the oligosaccharides linked by N-glycosylation, can account for a significant fraction of the surface area of the protein and can therefore be detrimental to crystallization. The preparation of recombinant proteins in *E. coli* eliminates these post-translational modifications and may sometimes solve the problem (Mohanty *et al.*, 2009), but N-glycosylation is often required for appropriate folding and solubility, so this approach is not always possible. However, if a eukaryotic expression system is a necessity, the problem can often be resolved by mutating the asparagines within the relevant glycosylation motifs (Asn-*X*-Thr/Ser), *e.g.* to aspartates, as was performed in the case of the extracellular domain of the metabotropic glutamate receptor expressed in insect cells (Muto *et al.*, 2009), or to glutamines, as was performed for the human testis angiotensin-converting enzyme (Gordon *et al.*, 2003). Alternatively, glycosylation at these sites can be eliminated by mutation of the Thr/Ser residues in the glycosylation motif to alanine or other amino acids, as described for rat cathepsin B (Lee *et al.*, 1990), or valine, as was the case with the Ebola virus glycoprotein (Lee *et al.*, 2008, 2009). Similarly, potentially glycosylated threonines or serines in O-glycosylated glycoproteins can be mutated to other amino acids (Horan *et al.*, 1998) to avoid or reduce glycosylation.

Other post-translational modifications occur less frequently. Prenylation and N-myristoylation can occur at the C- and N-termini, respectively. Expression in *E. coli*, often using truncated versions of target proteins, is a common remedy (Pai *et al.*, 1990).

4.3.8. Stabilization of protein targets

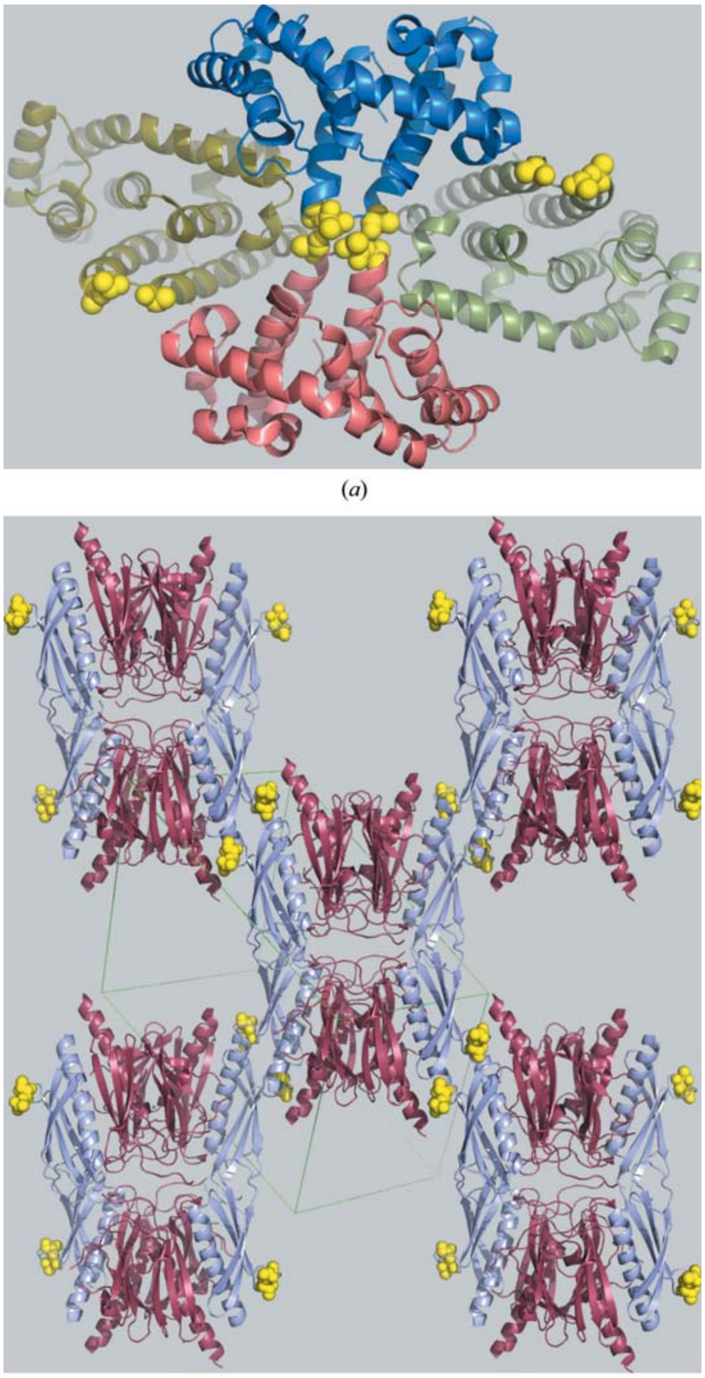
There is currently no clear consensus regarding a possible correlation between the thermostability of a protein and its propensity to form crystals. It is often assumed on the basis of somewhat anecdotal evidence that thermostable proteins are more readily crystallizable and therefore if a specific target protein is recalcitrant to crystallization then a homologue from a thermophilic organism should instead be used. In some cases, low thermostability may correlate with the presence of unstructured loops or termini and consequently the construct-optimization strategies, as described above, are likely to yield a more crystallizable variant with a concomitantly increased stability. For example, a study of MAPKAP kinase 2 showed that truncated variants with increased thermal stability also showed higher crystallization propensity (Malawski *et al.*, 2006). However, it is uncertain whether the relationship is causal or serendipitous. A recent analysis of large-scale data from a structural genomics project showed that when partly or fully unfolded proteins and hyperstable proteins (with melting temperatures T_m of greater than 90 °C) are excluded from comparisons, thermostability *per se* does not correlate with propensity for crystallization (Price *et al.*, 2009). Consequently, it appears that in a general case of a well behaving protein, attempts to increase thermostability by site-directed mutagenesis may not necessarily yield variants with enhanced crystallization properties, although when the prospective crystallization target is inherently unstable, engineering more stable variants may be helpful. In fact, this strategy has been successfully used for membrane proteins, which are often unstable in detergent environments. The first structure of a recombinant G-protein-coupled receptor (GPCR), *i.e.* bovine rhodopsin in complex with 11-*cis* retinal, was obtained using a thermostable variant with an engineered disulfide bond (Standfuss *et al.*, 2007). In the more recent case of the turkey adrenergic $\beta 2$ receptor, 318 variants were screened and six mutations were identified that increased thermostability. A variant containing all six mutations had an apparent T_m that was 21 °C higher than that of the native protein in dodecylmaltoside (DDM), was more stable in short-chain detergents and was successfully crystallized (Warne *et al.*, 2008, 2009). The effort required for such vast screening is substantial, but it appears that within protein families (such as GPCRs) the pattern of mutations enhancing thermostability is preserved, thus making it possible to transfer the mutations from one family member to another (Serrano-Vega & Tate, 2009).

Finally, it should be noted that protein thermostability is strongly dependent on solution parameters such as the ionic strength. In the case of ribonuclease SA, the melting temperature (T_m) increased from ~40 to ~60 °C on transfer from pure 50 m*M* diglycine buffer to 0.9 *M* ammonium sulfate (Trevino *et al.*, 2007). Similarly, binding small molecules, either physiological or non-physiological ligands, typically promote stability (Matulis *et al.*, 2005). High-throughput screening methods have been developed to aid in screening for conditions and ligands that enhance stability (Vedadi *et al.*, 2006; Mezzasalma *et al.*, 2007) and in a general case this strategy appears to hold better promise than attempts to engineer higher stability through mutagenesis.

4.3.9. Surface-entropy reduction (SER)

For well ordered intrinsically stable proteins that show none of the problems addressed above, the propensity of the molecules to associate together and form crystals mediated by weak but specific interactions is ultimately defined by the physical chemistry and topology of the molecular surface. As already pointed out, large flexible amino acids on the surface, such as Lys, Glu and Gln, constitute an impediment to intermolecular interactions and consequently to protein crystallization (Cieślik & Derewenda, 2009; Price *et al.*, 2009). In fact, it has been suggested that these residues and the 'entropy shield' that they form play a role in protein evolution which, given the high average concentration of proteins in cells, disfavours protein–protein interactions unless they are biologically functional (Doye, 2004). Thus, an intuitively obvious way to generate crystallizable variants is to replace selected large and surface-exposed residues with smaller residues such as alanine. This crystal-engineering strategy based on the surface-entropy reduction (SER) concept was extensively tested using as a model system the globular domain of the human Rho-specific guanine nucleotide-dissociation inhibitor (RhoGDI), which is recalcitrant to crystallization in its wild-type form owing to a high content of Lys and Glu residues, which constitute more than 20% of the sequence (Longenecker, Garrard *et al.*, 2001; Mateja *et al.*, 2002; Derewenda, 2004; Cooper *et al.*, 2007). These experiments established that in order to be most effective the SER strategy requires simultaneous mutations of clusters of two to three solvent-exposed high-entropy amino acids, typically Lys, Glu or Gln, located in close sequence proximity. These amino acids are replaced with alanine, although threonine and tyrosine, which is known to make a positive contribution at protein–protein interfaces such as antibody–antigen complexes (Fellouse *et al.*, 2006), can also be used (Cooper *et al.*, 2007). Engineered low-surface-entropy variants of RhoGDI produced new and unique crystal forms, many with superior diffraction quality when compared with the wild-type protein. Importantly, in the vast majority of these crystals the mutated surface patches mediated crystal contacts, suggesting that SER engineering directly drives crystallization in a rational fashion by creating suitable crystal-contact-forming interfaces. The general utility of the method was further established by the crystallization of several novel protein targets found to be recalcitrant to crystallization in their wild-type form (Longenecker, Lewis *et al.*, 2001; Derewenda *et al.*, 2004; Devedjiev *et al.*, 2004; Janda *et al.*, 2004).

SER is quickly becoming a method of choice for engineering crystallizable variants of both individual proteins and protein–protein complexes. To date (December 2009), there have already been more than 100 depositions made to the Protein Data Bank (Berman *et al.*, 2007) based on diffraction studies of crystals generated by SER and corresponding to 47 novel structures, seven novel protein–protein complexes, several studies of proteins in complexes with drug leads aimed at rational drug development and two membrane proteins. The current list of crystal structures obtained using SER crystals includes a number of cases of exceptionally high biological interest. For example, the EscJ protein from enteropathogenic *E. coli*, the oligomerization of which initiates assembly of the type III bacterial secretion system, was crystallized with three entropy-reducing mutations (E62A, K63A and E64A) forming a key contact in the crystal structure (Yip *et al.*, 2005). Likewise, the HIV CcmK4 capsid



(a)

(b)

Figure 4.3.9.1
Two examples of proteins crystallized by the surface-entropy reduction (SER) method. (*a*) The RGSL domain of the PDZRhoGEF nucleotide-exchange factor (PDB code 1htj; Longenecker, Lewis *et al.*, 2001); the yellow spheres show the alanines introduced by mutagenesis, which mediate an isologous crystal contact across a crystallographic twofold axis. (*b*) The crystal structure of EpsI complexed with EpsJ (PDB code 2ret; Yanez *et al.*, 2008); the EpsI protein (pale blue) contains two surface mutations, shown by yellow spheres, which mediate heterologous crystal contacts.

protein only crystallized after an entropy-reducing mutation (E104Y) was introduced into the protein (Pornillos *et al.*, 2009). The protein–protein complexes solved to date underscore the utility of the method, which extends beyond individual proteins because high-entropy patches occur outside complex interfaces. For example, the complex of c-Src and its inactivator Csk was crystallized using a variant of Csk carrying K361A and K362A mutations (Levinson *et al.*, 2008). Similarly, it was possible to crystallize the complex of two pseudopilins EpsI and EpsJ from the type 2 secretion system of *Vibrio vulnificus* when a variant of EpsI carrying two mutations (E128T and K129T) was used (Yanez *et al.*, 2008; Fig. 4.3.9.1).

An interesting variation of the SER method was used in the investigation of the RACK1 protein, which was crystallized as an in-line fusion with an MBP variant carrying D82A, K83A and K239A mutations (Ullah *et al.*, 2008). This is the first example of the application of the surface-entropy reduction strategy to a carrier protein and not the crystallization target itself.

The SER strategy is attractive not only because of its efficacy but also because of its simplicity: once an expression construct for a target protein is available several rounds of mutagenesis can easily create variants with systematically enhanced crystallizability. To assist in the design of crystallizable variants, a server has been developed that uses the amino-acid sequence of the target to identify suitable mutation sites (Goldschmidt *et al.*, 2007).

4.3.10. Improvement of crystal quality

In most cases, protein engineering is used as a tool of last resort to obtain variants for proteins for which no crystals can be grown using the wild-type form. However, it may sometimes be necessary to obtain a new, different crystal form even when the wild-type protein does crystallize. Such a need may arise, for example, in drug-design investigations, where high-resolution structures are particularly critical for evaluation of the interactions between lead compounds and the target protein and may not always be possible using wild-type crystals. A novel crystal form may also be necessary if the wild-type crystals contain the target protein in an orientation in which the active site is obscured by crystal contacts, making it impossible to soak in drug lead compounds and screen small-molecule libraries by high-throughput crystallography (Blundell & Patel, 2004).

One possible strategy for obtaining a new crystal form is to modify the existing crystal contacts by replacing some of the participating amino acids. While this approach occasionally leads to improvement of the X-ray data resolution (Liu *et al.*, 2007; Mizutani *et al.*, 2008), modification of crystal contacts is typically counterproductive as it abolishes the propensity of the target to crystallize in one form but does not necessarily induce another (Charron *et al.*, 2002). A more successful strategy is to generate a novel crystal form by engineering new crystal contacts through SER. For example, a novel crystal form of the insulin-like growth factor 1 receptor kinase domain, a putative drug target, was obtained using a double mutant (E1067A and E1069A); the new form diffracted to 1.5 Å resolution, whereas the wild-type crystals only diffracted to 2.7 Å resolution (Munshi *et al.*, 2003). In the case of the catalytic domain of activated factor XI, a key enzyme in the blood coagulation cascade and another potential drug target, a single K437A mutation allowed the preparation of a crystal form that diffracted to 2.0 Å resolution (Jin *et al.*, 2005). Entropy-reducing mutations were also key in the preparation of a crystal form of HIV-1 reverse transcriptase for structure-based drug design that diffracted to 1.8 Å resolution, in contrast to the typical 2.5–3.0 Å range observed for the wild-type protein crystals (Bauman *et al.*, 2008).

4.3.11. Conclusions

Protein engineering has become a routine tool that is used to generate crystallizable macromolecules and their complexes. While some approaches may only apply to very specific targets, a number of strategies offer general applicability. Among these, gene-construct optimization or surface-entropy reduction are quickly gaining popularity as methods of choice. However, it should be stressed that none of the methods described here offer a guarantee that the target protein can be coerced to crystallize. To maximize the chances of success, one must frequently attack the problem on multiple fronts based on an understanding of the chemical and physical properties of a specific protein. This is particularly true of technically difficult targets such as membrane proteins. A classic example illustrating this principle is the study of the HIV gp120 envelope glycoprotein (Kwong *et al.*, 1998, 1999). The construct that was ultimately used in successful crystallization screens had deletions of 52 and 19 residues from the N- and C-termini and two flexible loops replaced by Gly-Ala-Gly linkages; additionally, the protein was 90% deglycosylated compared with the wild type. Moreover, this engineered gp120 was only crystallized in the form of a ternary complex with the CD4 receptor and an Fab fragment from a neutralizing antibody. In the recent case of the ATP-gated P2X4 ion channel, a crystallizable variant was obtained after a series of N- and C-terminal deletions were screened to identify the smallest functional unit and the introduction of three mutations (C51F/N78K/N187R) to eliminate both aggregation arising from oxidation and N-glycosylation (Kawate *et al.*, 2009).

The rapidly expanding database of macromolecular structures greatly enhances our understanding of the physical chemistry of proteins, ultimately enhancing our ability to predict the behaviour of a protein in solution from its sequence. It is therefore increasingly possible to rely on such theoretical predictions *in lieu* of tedious experimental screens. A number of online tools have been developed for this purpose. The propensity of a protein target to crystallize can be evaluated using the *XtalPred* server (http://ffas.burnham.org/XtalPred-cgi/xtal.pl), which offers insights into potential sources of problems arising from sequence features (Slabinski *et al.*, 2007). Automated design of optimally truncated constructs for structural analysis has been made possible by the *ProteinCCD* meta-server (http://xtal.nki.nl/ccd), which uses the cDNA sequence of the target (Mooij *et al.*, 2009). This server collects information about secondary structure, disorder, putative coiled coils, transmembrane segments, domains and domain linkers, and suggests constructs so that the user can interactively choose suitable options and obtain sequences of oligonucleotides needed for appropriate PCR amplification (Mooij *et al.*, 2009). For proteins recalcitrant to crystallization in their wild-type form, surface mutations enhancing crystallizability can be designed using the surface-entropy reduction server (http://nihserver.mbi.ucla.edu/SER/; Goldschmidt *et al.*, 2007).

As the focus of macromolecular crystallography shifts from the principles of protein architecture to increasingly complex biological questions, the approach to crystallization is also undergoing dramatic evolution. As we gain better understanding of the microscopic nature of protein crystallization, we will be able to develop rational protein-engineering strategies that systematically and significantly improve the success rate of crystallization.

The research on protein crystallization in my laboratory is supported by the National Institute for General Medical Sciences. I am greatly indebted to Drs Urszula Derewenda, Anthony Kossiakoff, Alexander McPherson, Geoff Waldo and David Cooper for discussions and comments on the early version of the manuscript.

This chapter is based on a recent article published in *Acta Crystallographica Section D* (Derewenda, 2010).

References

Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H. R. & Iwata, S. (2003). *Structure and mechanism of the lactose permease of Escherichia coli. Science*, **301**, 610–615.

Al-Ayyoubi, M., Gettins, P. G. & Volz, K. (2004). *Crystal structure of human maspin, a serpin with antitumor properties: reactive center loop of maspin is exposed but constrained. J. Biol. Chem.* **279**, 55540–55544.

Aleshin, A. E., Hoffman, C., Firsov, L. M. & Honzatko, R. B. (1994). *Refined crystal structures of glucoamylase from Aspergillus awamori var. X100. J. Mol. Biol.* **238**, 575–591.

Bandeiras, T. M. *et al.* (2008). *Structure of wild-type Plk-1 kinase domain in complex with a selective DARPin. Acta Cryst.* D**64**, 339–353.

Bauman, J. D., Das, K., Ho, W. C., Baweja, M., Himmel, D. M., Clark, A. D. Jr, Oren, D. A., Boyer, P. L., Hughes, S. H., Shatkin, A. J. & Arnold, E. (2008). *Crystal engineering of HIV-1 reverse transcriptase for structure-based drug design. Nucleic Acids Res.* **36**, 5083–5092.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res.* **35**, D301–D303.

Birtley, J. R. & Curry, S. (2005). *Crystallization of foot-and-mouth disease virus 3C protease: surface mutagenesis and a novel crystal-optimization strategy. Acta Cryst.* D**61**, 646–650.

Blundell, T. L. & Patel, S. (2004). *High-throughput X-ray crystallography for drug discovery. Curr. Opin. Pharmacol.* **4**, 490–496.

Brewer, S. J., Haymore, B. L., Hopp, T. P. & Sassenfeld, H. M. (1991). *Engineering proteins to enable their isolation in a biologically active form. Bioprocess Technol.* **12**, 239–266.

Brooun, A., Foster, S. A., Chrencik, J. E., Chien, E. Y., Kolatkar, A. R., Streiff, M., Ramage, P., Widmer, H., Weckbecker, G. & Kuhn, P. (2007). *Remedial strategies in structural proteomics: expression, purification, and crystallization of the Vav1/Rac1 complex. Protein Expr. Purif.* **53**, 51–62.

Cabantous, S., Pedelacq, J. D., Mark, B. L., Naranjo, C., Terwilliger, T. C. & Waldo, G. S. (2005). *Recent advances in GFP folding reporter and split-GFP solubility reporter technologies. Application to improving the folding and solubility of recalcitrant proteins from Mycobacterium tuberculosis. J. Struct. Funct. Genomics*, **6**, 113–119.

Cabantous, S. & Waldo, G. S. (2006). *In vivo and in vitro protein solubility assays using split GFP. Nat. Methods*, **3**, 845–854.

Campbell, J. W., Duée, E., Hodgson, G., Mercer, W. D., Stammers, D. K., Wendell, P. L., Muirhead, H. & Watson, H. C. (1972). *X-ray diffraction studies on enzymes in the glycolytic pathway. Cold Spring Harb. Symp. Quant. Biol.* **36**, 165–170.

Canaves, J. M., Page, R., Wilson, I. A. & Stevens, R. C. (2004). *Protein biophysical properties that correlate with crystallization success in Thermotoga maritima: maximum clustering strategy for structural genomics. J. Mol. Biol.* **344**, 977–991.

Carter, C. W. Jr & Carter, C. W. (1979). *Protein crystallization using incomplete factorial experiments. J. Biol. Chem.* **254**, 12219–12223.

Carugo, O. & Argos, P. (1997). *Protein–protein crystal-packing contacts. Protein Sci.* **6**, 2261–2263.

Center, R. J., Kobe, B., Wilson, K. A., Teh, T., Howlett, G. J., Kemp, B. E. & Poumbourios, P. (1998). *Crystallization of a trimeric human T cell leukemia virus type 1 gp21 ectodomain fragment as a chimera with maltose-binding protein. Protein Sci.* **7**, 1612–1619.

Charron, C., Kern, D. & Giegé, R. (2002). *Crystal contacts engineering of aspartyl-tRNA synthetase from Thermus thermophilus: effects on crystallizability. Acta Cryst.* D**58**, 1729–1733.

Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K. & Stevens, R. C. (2007). *High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor. Science*, **318**, 1258–1265.

Cieślik, M. & Derewenda, Z. S. (2009). *The role of entropy and polarity in intermolecular contacts in protein crystals. Acta Cryst.* D**65**, 500–509.

Cohen, S. L., Ferré-D'Amaré, A. R., Burley, S. K. & Chait, B. T. (1995). *Probing the solution structure of the DNA-binding protein max by a combination of proteolysis and mass-spectrometry. Protein Sci.* **4**, 1088–1099.

Cooper, D. R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M. & Derewenda, Z. (2007). *Protein crystallization by* *surface entropy reduction: optimization of the SER strategy*. *Acta Cryst.* D**63**, 636–645.

Corsini, L., Hothorn, M., Scheffzek, K., Sattler, M. & Stier, G. (2008). *Thioredoxin as a fusion tag for carrier-driven crystallization*. *Protein Sci.* **17**, 2070–2079.

Dale, G. E., Broger, C., Langen, H., D'Arcy, A. & Stuber, D. (1994). *Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase*. *Protein Eng.* **7**, 933–939.

Derewenda, U., Mateja, A., Devedjiev, Y., Routzahn, K. M., Evdokimov, A. G., Derewenda, Z. S. & Waugh, D. S. (2004). *The structure of Yersinia pestis V-antigen, an essential virulence factor and mediator of immunity against plague*. *Structure*, **12**, 301–306.

Derewenda, Z. S. (2004). *Rational protein crystallization by mutational surface engineering*. *Structure*, **12**, 529–535.

Derewenda, Z. S. (2010). *Application of protein engineering to enhance crystallizability and improve crystal properties*. *Acta Cryst.* D**66**, 604–615.

Derewenda, Z. S. & Vekilov, P. G. (2006). *Entropy and surface engineering in protein crystallization*. *Acta Cryst.* D**62**, 116–124.

Devedjiev, Y., Surendranath, Y., Derewenda, U., Gabrys, A., Cooper, D. R., Zhang, R. G., Lezondra, L., Joachimiak, A. & Derewenda, Z. S. (2004). *The structure and ligand binding properties of the B. subtilis YkoF gene product, a member of a novel family of thiamin/HMP-binding proteins*. *J. Mol. Biol.* **343**, 395–406.

Dong, A. *et al.* (2007). *In situ proteolysis for protein crystallization and structure determination*. *Nat. Methods*, **4**, 1019–1021.

Doye, J. P. K. (2004). *Inhibition of protein crystallization by evolutionary negative design*. *Phys. Biol.* **1**, P9–P13.

Dutzler, R., Campbell, E. B. & MacKinnon, R. (2003). *Gating the selectivity filter in ClC chloride channels*. *Science*, **300**, 108–112.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R. & Davies, D. R. (1994). *Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases*. *Science*, **266**, 1981–1986.

Eichinger, A., Nasreen, A., Kim, H. J. & Skerra, A. (2007). *Structural insight into the dual ligand specificity and mode of high density lipoprotein association of apolipoprotein D*. *J. Biol. Chem.* **282**, 31068–31075.

Engel, C. K., Chen, L. & Privé, G. G. (2002). *Insertion of carrier proteins into hydrophilic loops of the Escherichia coli lactose permease*. *Biochim. Biophys. Acta*, **1564**, 38–46.

Erdemir, D., Lee, A. Y. & Myerson, A. S. (2009). *Nucleation of crystals from solution: classical and two-step models*. *Acc. Chem. Res.* **42**, 621–629.

Farinas, E. T. (2006). *Directed evolution approaches for protein engineering*. *Comb. Chem. High Throughput Screen.* **9**, 235–236.

Farinas, E. T., Bulter, T. & Arnold, F. H. (2001). *Directed enzyme evolution*. *Curr. Opin. Biotechnol.* **12**, 545–551.

Fellouse, F. A., Barthelemy, P. A., Kelley, R. F. & Sidhu, S. S. (2006). *Tyrosine plays a dominant functional role in the paratope of a synthetic antibody derived from a four amino acid code*. *J. Mol. Biol.* **357**, 100–114.

Fellouse, F. A., Wiesmann, C. & Sidhu, S. S. (2004). *Synthetic antibodies from a four-amino-acid code: a dominant role for tyrosine in antigen recognition*. *Proc. Natl Acad. Sci. USA*, **101**, 12467–12472.

Finkelstein, A. V. & Janin, J. (1989). *The price of lost freedom: entropy of bimolecular complex formation*. *Protein Eng.* **3**, 1–3.

Georgalis, Y., Umbach, P., Raptis, J. & Saenger, W. (1997). *Lysozyme aggregation studied by light scattering. I. Influence of concentration and nature of electrolytes*. *Acta Cryst.* D**53**, 691–702.

Gilbreth, R. N., Esaki, K., Koide, A., Sidhu, S. S. & Koide, S. (2008). *A dominant conformational role for amino acid diversity in minimalist protein–protein interfaces*. *J. Mol. Biol.* **381**, 407–418.

Gliko, O., Neumaier, N., Pan, W., Haase, I., Fischer, M., Bacher, A., Weinkauf, S. & Vekilov, P. G. (2005). *A metastable prerequisite for the growth of lumazine synthase crystals*. *J. Am. Chem. Soc.* **127**, 3433–3438.

Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K. & Riggs, A. D. (1979). *Expression in Escherichia coli of chemically synthesized genes for human insulin*. *Proc. Natl Acad. Sci. USA*, **76**, 106–110.

Goldschmidt, L., Cooper, D. R., Derewenda, Z. S. & Eisenberg, D. (2007). *Toward rational protein crystallization: A web server for the design of crystallizable protein variants*. *Protein Sci.* **16**, 1569–1576.

Gordon, K., Redelinghuys, P., Schwager, S. L., Ehlers, M. R., Papageorgiou, A. C., Natesh, R., Acharya, K. R. & Sturrock, E. D. (2003). *Deglycosylation, processing and crystallization of human testis angiotensin-converting enzyme*. *Biochem. J.* **371**, 437–442.

Hamuro, Y., Coales, S. J., Southern, M. R., Nemeth-Cawley, J. F., Stranz, D. D. & Griffin, P. R. (2003). *Rapid analysis of protein structure and dynamics by hydrogen/deuterium exchange mass spectrometry*. *J. Biomol. Tech.* **14**, 171–182.

Hawkins, R. E., Russell, S. J. & Winter, G. (1992). *Selection of phage antibodies by binding-affinity – mimicking affinity maturation*. *J. Mol. Biol.* **226**, 889–896.

He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N. & Dunker, A. K. (2009). *Predicting intrinsic disorder in proteins: an overview*. *Cell Res.* **19**, 929–949.

Horan, T. P., Simonet, L., Jacobsen, R., Mann, M., Haniu, M., Wen, J., Arakawa, T., Kuwamoto, M. & Martin, F. (1998). *Coexpression of G-CSF with an unglycosylated G-CSF receptor mutant results in secretion of a stable complex*. *Protein Expr. Purif.* **14**, 45–53.

Hunte, C. & Michel, H. (2002). *Crystallisation of membrane proteins mediated by antibody fragments*. *Curr. Opin. Struct. Biol.* **12**, 503–508.

Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F. & Boyer, H. W. (1977). *Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin*. *Science*, **198**, 1056–1063.

Jancarik, J. & Kim, S.-H. (1991). *Sparse matrix sampling: a screening method for crystallization of proteins*. *J. Appl. Cryst.* **24**, 409–411.

Janda, I., Devedjiev, Y., Derewenda, U., Dauter, Z., Bielnicki, J., Cooper, D. R., Graf, P. C., Joachimiak, A., Jakob, U. & Derewenda, Z. S. (2004). *The crystal structure of the reduced, Zn2+-bound form of the B. subtilis Hsp33 chaperone and its implications for the activation mechanism*. *Structure*, **12**, 1901–1907.

Janin, J. (1997). *Specific versus non-specific contacts in protein crystals*. *Nat. Struct. Biol.* **4**, 973–974.

Janin, J. & Rodier, F. (1995). *Protein–protein interaction at crystal contacts*. *Proteins*, **23**, 580–587.

Jenkins, T. M., Hickman, A. B., Dyda, F., Ghirlando, R., Davies, D. R. & Craigie, R. (1995). *Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues*. *Proc. Natl Acad. Sci. USA*, **92**, 6057–6061.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T. & MacKinnon, R. (2003). *X-ray structure of a voltage-dependent K+ channel*. *Nature (London)*, **423**, 33–41.

Jin, L., Pandey, P., Babine, R. E., Weaver, D. T., Abdel-Meguid, S. S. & Strickler, J. E. (2005). *Mutation of surface residues to promote crystallization of activated factor XI as a complex with benzamidine: an essential step for the iterative structure-based design of factor XI inhibitors*. *Acta Cryst.* D**61**, 1418–1425.

Karpusas, M., Lucci, J., Ferrant, J., Benjamin, C., Taylor, F. R., Strauch, K., Garber, E. & Hsu, Y. M. (2001). *Structure of CD40 ligand in complex with the Fab fragment of a neutralizing humanized antibody*. *Structure*, **9**, 321–329.

Kawate, T., Michel, J. C., Birdsong, W. T. & Gouaux, E. (2009). *Crystal structure of the ATP-gated P2X(4) ion channel in the closed state*. *Nature (London)*, **460**, 592–598.

Ke, A. & Wolberger, C. (2003). *Insights into binding cooperativity of MATa1/MATalpha2 from the crystal structure of a MATa1 homeodomain-maltose binding protein chimera*. *Protein Sci.* **12**, 306–312.

Kendrew, J. C., Parrish, R. G., Marrack, J. R. & Orlans, E. S. (1954). *The species specificity of myoglobin*. *Nature (London)*, **174**, 946–949.

Koide, A., Gilbreth, R. N., Esaki, K., Tereshko, V. & Koide, S. (2007). *High-affinity single-domain binding proteins with a binary-code interface*. *Proc. Natl Acad. Sci. USA*, **104**, 6632–6637.

Koide, A., Tereshko, V., Uysal, S., Margalef, K., Kossiakoff, A. A. & Koide, S. (2007). *Exploring the capacity of minimalist protein interfaces: interface energetics and affinity maturation to picomolar K-D of a single-domain antibody with a flat paratope*. *J. Mol. Biol.* **373**, 941–953.

Koide, S. (2009). *Engineering of recombinant crystallization chaperones*. *Curr. Opin. Struct. Biol.* **19**, 449–457.

Korotkov, K. V., Pardon, E., Steyaert, J. & Hol, W. G. (2009). *Crystal structure of the N-terminal domain of the secretin GspD from ETEC determined with the assistance of a nanobody*. *Structure*, **17**, 255–265.

Kovari, L. C., Momany, C. & Rossmann, M. G. (1995). *The use of antibody fragments for crystallization and structure determinations*. *Structure*, **3**, 1291–1293.

Kuge, M., Fujii, Y., Shimizu, T., Hirose, F., Matsukage, A. & Hakoshima, T. (1997). *Use of a fusion protein to obtain crystals suitable for X-ray analysis: crystallization of a GST-fused protein containing the DNA-binding domain of DNA replication-related element-binding factor, DREF*. *Protein Sci.* **6**, 1783–1786.

Kwong, P. D., Wyatt, R., Desjardins, E., Robinson, J., Culp, J. S., Hellmig, B. D., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1999). *Probability analysis of variational crystallization and its application to gp120, the exterior envelope glycoprotein of type 1 human immunodeficiency virus (HIV-1)*. *J. Biol. Chem.* **274**, 4115–4123.

Kwong, P. D., Wyatt, R., Robinson, J., Sweet, R. W., Sodroski, J. & Hendrickson, W. A. (1998). *Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody*. *Nature (London)*, **393**, 648–659.

Lam, A. Y., Pardon, E., Korotkov, K. V., Hol, W. G. & Steyaert, J. (2009). *Nanobody-aided structure determination of the EpsI:EpsJ pseudopilin heterodimer from Vibrio vulnificus*. *J. Struct. Biol.* **166**, 8–15.

Lawson, D. M., Artymiuk, P. J., Yewdall, S. J., Smith, J. M., Livingstone, J. C., Treffry, A., Luzzago, A., Levi, S., Arosio, P., Cesareni, G., Thomas, C. D., Shaw, W. V. & Harrison, P. M. (1991). *Solving the structure of human H ferritin by genetically engineering intermolecular crystal contacts*. *Nature (London)*, **349**, 541–544.

Lee, C. V., Liang, W. C., Dennis, M. S., Eigenbrot, C., Sidhu, S. S. & Fuh, G. (2004). *High-affinity human antibodies from phage-displayed synthetic Fab libraries with a single framework scaffold*. *J. Mol. Biol.* **340**, 1073–1093.

Lee, J. E., Fusco, M. L., Abelson, D. M., Hessell, A. J., Burton, D. R. & Saphire, E. O. (2009). *Techniques and tactics used in determining the structure of the trimeric ebolavirus glycoprotein*. *Acta Cryst.* D**65**, 1162–1180.

Lee, J. E., Fusco, M. L., Hessell, A. J., Oswald, W. B., Burton, D. R. & Saphire, E. O. (2008). *Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor*. *Nature (London)*, **454**, 177–182.

Lee, S. Y., Lee, A., Chen, J. & MacKinnon, R. (2005). *Structure of the KvAP voltage-dependent K+ channel and its dependence on the lipid membrane*. *Proc. Natl Acad. Sci. USA*, **102**, 15441–15446.

Lee, X., Ahmed, F. R., Hirama, T., Huber, C. P., Rose, D. R., To, R., Hasnain, S., Tam, A. & Mort, J. S. (1990). *Crystallization of recombinant rat cathepsin B*. *J. Biol. Chem.* **265**, 5950–5951.

Levinson, N. M., Seeliger, M. A., Cole, P. A. & Kuriyan, J. (2008). *Structural basis for the recognition of c-Src by its inactivator Csk*. *Cell*, **134**, 124–134.

Lipovsek, D. & Pluckthun, A. (2004). *In-vitro protein evolution by ribosome display and mRNA display*. *J. Immunol. Methods*, **290**, 51–67.

Liu, B., Luna, V. M., Chen, Y., Stout, C. D. & Fee, J. A. (2007). *An unexpected outcome of surface engineering an integral membrane protein: improved crystallization of cytochrome ba_3 from Thermus thermophilus*. *Acta Cryst.* F**63**, 1029–1034.

Longenecker, K. L., Garrard, S. M., Sheffield, P. J. & Derewenda, Z. S. (2001). *Protein crystallization by rational mutagenesis of surface residues: Lys to Ala mutations promote crystallization of RhoGDI*. *Acta Cryst.* D**57**, 679–688.

Longenecker, K. L., Lewis, M. E., Chikumi, H., Gutkind, J. S. & Derewenda, Z. S. (2001). *Structure of the RGS-like domain from PDZ-RhoGEF: linking heterotrimeric G protein-coupled signaling to Rho GTPases*. *Structure*, **9**, 559–569.

McElroy, H. H., Sisson, G. W., Schottlin, W. E., Aust, R. M. & Villafranca, J. E. (1992). *Studies on engineering crystallizability by mutation of surface residues of human thymidylate synthase*. *J. Cryst. Growth*, **122**, 265–272.

McPherson, A. (1982). *The Preparation and Analysis of Protein Crystals*. New York: John Wiley & Sons.

Malawski, G. A., Hillig, R. C., Monteclaro, F., Eberspaecher, U., Schmitz, A. A., Crusius, K., Huber, M., Egner, U., Donner, P. & Muller-Tiemann, B. (2006). *Identifying protein construct variants with increased crystallization propensity – a case study*. *Protein Sci.* **15**, 2718–2728.

Malhotra, A. (2009). *Tagging for protein expression*. *Methods Enzymol.* **463**, 239–258.

Mark, B. L., Mahuran, D. J., Cherney, M. M., Zhao, D., Knapp, S. & James, M. N. (2003). *Crystal structure of human beta-hexosaminidase B: understanding the molecular basis of Sandhoff and Tay-Sachs disease*. *J. Mol. Biol.* **327**, 1093–1109.

Mateja, A., Devedjiev, Y., Krowarsch, D., Longenecker, K., Dauter, Z., Otlewski, J. & Derewenda, Z. S. (2002). *The impact of Glu→Ala and Glu→Asp mutations on the crystallization properties of RhoGDI: the structure of RhoGDI at 1.3 Å resolution*. *Acta Cryst.* D**58**, 1983–1991.

Matulis, D., Kranz, J. K., Salemme, F. R. & Todd, M. J. (2005). *Thermodynamic stability of carbonic anhydrase: measurements of binding affinity and stoichiometry using ThermoFluor*. *Biochemistry*, **44**, 5258–5266.

Mezzasalma, T. M., Kranz, J. K., Chan, W., Struble, G. T., Schalk-Hihi, C., Deckman, I. C., Springer, B. A. & Todd, M. J. (2007). *Enhancing recombinant protein quality and yield by protein stability profiling*. *J. Biomol. Screen.* **12**, 418–428.

Mizutani, H., Saraboji, K., Malathy Sony, S. M., Ponnuswamy, M. N., Kumarevel, T., Krishna Swamy, B. S., Simanshu, D. K., Murthy, M. R. N. & Kunishima, N. (2008). *Systematic study on crystal-contact engineering of diphthine synthase: influence of mutations at crystal-packing regions on X-ray diffraction quality*. *Acta Cryst.* D**64**, 1020–1033.

Mohanty, A. K., Fisher, A. J., Yu, Z., Pradeep, M. A., Janjanam, J. & Kaushik, J. K. (2009). *Cloning, expression, characterization and crystallization of BRP39, a signalling glycoprotein expressed during mammary gland apoptosis*. *Protein Expr. Purif.* **64**, 213–218.

Monne, M., Han, L., Schwend, T., Burendahl, S. & Jovine, L. (2008). *Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats*. *Nature (London)*, **456**, 653–657.

Mooij, W. T. M., Mitsiki, E. & Perrakis, A. (2009). *ProteinCCD: enabling the design of protein truncation constructs for expression and crystallization experiments*. *Nucleic Acids Res.* **37**, W402–W405.

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. & Erlich, H. (1986). *Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction*. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273.

Munshi, S., Hall, D. L., Kornienko, M., Darke, P. L. & Kuo, L. C. (2003). *Structure of apo, unactivated insulin-like growth factor-1 receptor kinase at 1.5 Å resolution*. *Acta Cryst.* D**59**, 1725–1730.

Muto, T., Tsuchiya, D., Morikawa, K. & Jingami, H. (2009). *Site-specific unglycosylation to improve crystallization of the metabotropic glutamate receptor 3 extracellular domain*. *Acta Cryst.* F**65**, 236–241.

Nasreen, A., Vogt, M., Kim, H. J., Eichinger, A. & Skerra, A. (2006). *Solubility engineering and crystallization of human apolipoprotein D*. *Protein Sci.* **15**, 190–199.

Nauli, S., Farr, S., Lee, Y.-J., Kim, H.-Y., Faham, S. & Bowie, J. U. (2007). *Polymer-driven crystallization*. *Protein Sci.* **16**, 2542–2551.

Neau, D. B., Gilbert, N. C., Bartlett, S. G., Dassey, A. & Newcomer, M. E. (2007). *Improving protein crystal quality by selective removal of a Ca^{2+}-dependent membrane-insertion loop*. *Acta Cryst.* F**63**, 972–975.

Nettleship, J. E., Ren, J., Rahman, N., Berrow, N. S., Hatherley, D., Barclay, A. N. & Owens, R. J. (2008). *A pipeline for the production of antibody fragments for structural studies using transient expression in HEK 293T cells*. *Protein Expr. Purif.* **62**, 83–89.

Niemann, H. H., Schmoldt, H. U., Wentzel, A., Kolmar, H. & Heinz, D. W. (2006). *Barnase fusion as a tool to determine the crystal structure of the small disulfide-rich protein McoEeTI*. *J. Mol. Biol.* **356**, 1–8.

Niessing, D., Huttelmaier, S., Zenklusen, D., Singer, R. H. & Burley, S. K. (2004). *She2p is a novel RNA binding protein with a basic helical hairpin motif*. *Cell*, **119**, 491–502.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. & Dunker, A. K. (2003). *Predicting intrinsic disorder from amino acid sequence*. *Proteins*, **53**, 566–572.

Ostermeier, C., Iwata, S., Ludwig, B. & Michel, H. (1995). *F-V fragment mediated crystallization of the membrane-protein bacterial cytochrome-c-oxidase*. *Nat. Struct. Biol.* **2**, 842–846.

Pai, E. F., Krengel, U., Petsko, G. A., Goody, R. S., Kabsch, W. & Wittinghofer, A. (1990). *Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis*. *EMBO J.* **9**, 2351–2359.

Pantazatos, D., Kim, J. S., Klock, H. E., Stevens, R. C., Wilson, I. A., Lesley, S. A. & Woods, V. L. Jr (2004). *Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS*. *Proc. Natl Acad. Sci. USA*, **101**, 751–756.

Patel, S. B., Cameron, P. M., Frantz-Wattley, B., O'Neill, E., Becker, J. W. & Scapin, G. (2004). *Lattice stabilization and enhanced diffraction in human p38 alpha crystals by protein engineering*. *Biochim. Biophys. Acta*, **1696**, 67–73.

Peabody, D. S. & Al-Bitar, L. (2001). *Isolation of viral coat protein mutants with altered assembly and aggregation properties*. *Nucleic Acids Res.* **29**, E113.

Pédelacq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C.-Y., Rho, B.-S., Park, M. S., Terwilliger, T. C. & Waldo, G. S. (2002). *Engineering soluble proteins for structural genomics*. *Nat. Biotechnol.* **20**, 927–932.

Pellicane, G., Smith, G. & Sarkisov, L. (2008). *Molecular dynamics characterization of protein crystal contacts in aqueous solutions*. *Phys. Rev. Lett.* **101**, 248102.

Petsev, D. N., Thomas, B. R., Yau, S. T., Tsekova, D., Nanev, C., Wilson, W. W. & Vekilov, P. G. (2001). *Temperature-independent solubility and interactions between apoferritin monomers and dimers in solution*. *J. Cryst. Growth*, **232**, 21–29.

Pornillos, O., Ganser-Pornillos, B. K., Kelly, B. N., Hua, Y., Whitby, F. G., Stout, C. D., Sundquist, W. I., Hill, C. P. & Yeager, M. (2009). *X-ray structures of the hexameric building block of the HIV capsid*. *Cell*, **137**, 1282–1292.

Price, W. N. II *et al.* (2009). *Understanding the physical properties that control protein crystallization by analysis of large-scale experimental data*. *Nat. Biotechnol.* **27**, 51–57.

Privé, G. G., Verner, G. E., Weitzman, C., Zen, K. H., Eisenberg, D. & Kaback, H. R. (1994). *Fusion proteins as tools for crystallization: the lactose permease from Escherichia coli*. *Acta Cryst.* D**50**, 375–379.

Prongay, A. J., Smith, T. J., Rossmann, M. G., Ehrlich, L. S., Carter, C. A. & McClure, J. (1990). *Fusion proteins as tools for crystallization: the lactose permease from Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **87**, 9980–9984.

Robien, M. A., Nguyen, K. T., Kumar, A., Hirsh, I., Turley, S., Pei, D. & Hol, W. G. (2004). *An improved crystal form of Plasmodium falciparum peptide deformylase*. *Protein Sci.* **13**, 1155–1163.

Rosenbaum, D. M., Cherezov, V., Hanson, M. A., Rasmussen, S. G. F., Thian, F. S., Kobilka, T. S., Choi, H. J., Yao, X. J., Weis, W. I., Stevens, R. C. & Kobilka, B. K. (2007). *GPCR engineering yields high-resolution structural insights into beta(2)-adrenergic receptor function*. *Science*, **318**, 1266–1273.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). *Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase*. *Science*, **239**, 487–491.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. (1985). *Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia*. *Science*, **230**, 1350–1354.

Sassenfeld, H. M. (1990). *Engineering proteins for purification*. *Trends Biotechnol.* **8**, 88–93.

Schwartz, T. U., Walczak, R. & Blobel, G. (2004). *Circular permutation as a tool to reduce surface entropy triggers crystallization of the signal recognition particle receptor beta subunit*. *Protein Sci.* **13**, 2814–2818.

Sennhauser, G., Amstutz, P., Briand, C., Storchenegger, O. & Grütter, M. G. (2007). *Drug export pathway of multidrug exporter AcrB revealed by DARPin inhibitors*. *PLoS Biol.* **5**, 106–113.

Sennhauser, G. & Grütter, M. G. (2008). *Chaperone-assisted crystallography with DARPins*. *Structure*, **16**, 1443–1453.

Serrano-Vega, M. J. & Tate, C. G. (2009). *Transferability of thermostabilizing mutations between beta-adrenergic receptors*. *Mol. Membr. Biol.* **26**, 385–396.

Sharma, S., Zheng, H., Huang, Y. P. J., Ertekin, A., Hamuro, Y., Rossi, P., Tejero, R., Acton, T. B., Xiao, R., Jiang, M., Zhao, L., Ma, L.-C., Swapna, G. V. T., Aramini, J. M. & Montelione, G. T. (2009). *Construct optimization for protein NMR structure analysis using amide hydrogen/deuterium exchange mass spectrometry*. *Proteins*, **76**, 882–894.

Slabinski, L., Jaroszewski, L., Rychlewski, L., Wilson, I. A., Lesley, S. A. & Godzik, A. (2007). *XtalPred: a web server for prediction of protein crystallizability*. *Bioinformatics*, **23**, 3403–3405.

Smyth, D. R., Mrozkiewicz, M. K., McGrath, W. J., Listwan, P. & Kobe, B. (2003). *Crystal structures of fusion proteins with large-affinity tags*. *Protein Sci.* **12**, 1313–1322.

Standfuss, J., Xie, G., Edwards, P. C., Burghammer, M., Oprian, D. D. & Schertler, G. F. (2007). *Crystal structure of a thermally stable rhodopsin mutant*. *J. Mol. Biol.* **372**, 1179–1188.

Thornton, J. M. & Sibanda, B. L. (1983). *Amino and carboxy-terminal regions in globular-proteins*. *J. Mol. Biol.* **167**, 443–460.

Tidor, B. & Karplus, M. (1994). *The contribution of vibrational entropy to molecular association. The dimerization of insulin*. *J. Mol. Biol.* **238**, 405–414.

Trevino, S. R., Scholtz, J. M. & Pace, C. N. (2007). *Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa*. *J. Mol. Biol.* **366**, 449–460.

Trevino, S. R., Scholtz, J. M. & Pace, C. N. (2008). *Measuring and increasing protein solubility*. *J. Pharm. Sci.* **97**, 4155–4166.

Uhlen, M., Forsberg, G., Moks, T., Hartmanis, M. & Nilsson, B. (1992). *Fusion proteins in biotechnology*. *Curr. Opin. Biotechnol.* **3**, 363–369.

Ullah, H., Scappini, E. L., Moon, A. F., Williams, L. V., Armstrong, D. L. & Pedersen, L. C. (2008). *Structure of a signal transduction regulator, RACK1, from Arabidopsis thaliana*. *Protein Sci.* **17**, 1771–1780.

Uysal, S., Vasquez, V., Tereshko, V., Esaki, K., Fellouse, F. A., Sidhu, S. S., Koide, S., Perozo, E. & Kossiakoff, A. (2009). *Crystal structure of full-length KcsA in its closed conformation*. *Proc. Natl Acad. Sci. USA*, **106**, 6644–6649.

Vedadi, M. *et al.* (2006). *Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination*. *Proc. Natl Acad. Sci. USA*, **103**, 15835–15840.

Veesler, D., Dreier, B., Blangy, S., Lichière, J., Tremblay, D., Moineau, S., Spinelli, S., Tegoni, M., Plückthun, A., Campanacci, V. & Cambillau, C. (2009). *Crystal structure and function of a DARPin neutralizing inhibitor of lactococcal phage TP901-1: comparison of DARPin and camelid VHH binding mode*. *J. Biol. Chem.* **284**, 30718–30726.

Vekilov, P. G. (2003). *Solvent entropy effects in the formation of protein solid phases*. *Methods Enzymol.* **368**, 84–105.

Vekilov, P. G. (2004). *Dense liquid precursor for the nucleation of ordered solid phases from solution*. *Cryst. Growth Des.* **4**, 671–685.

Vekilov, P. G., Feeling-Taylor, A., Yau, S.-T. & Petsev, D. (2002). *Solvent entropy contribution to the free energy of protein crystallization*. *Acta Cryst.* D**58**, 1611–1616.

Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). *Flavors of protein disorder*. *Proteins*, **52**, 573–584.

Waldo, G. S. (2003). *Genetic screens and directed evolution for protein solubility*. *Curr. Opin. Chem. Biol.* **7**, 33–38.

Waldo, G. S., Standish, B. M., Berendzen, J. & Terwilliger, T. C. (1999). *Rapid protein-folding assay using green fluorescent protein*. *Nature Biotechnol.* **17**, 691–695.

Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G., Tate, C. G. & Schertler, G. F. (2008). *Structure of a beta1-adrenergic G-protein-coupled receptor*. *Nature (London)*, **454**, 486–491.

Warne, T., Serrano-Vega, M. J., Tate, C. G. & Schertler, G. F. (2009). *Development and crystallization of a minimal thermostabilized G protein-coupled receptor*. *Protein Expr. Purif.* **65**, 204–213.

Wernimont, A. & Edwards, A. (2009). *In situ proteolysis to generate crystals for structure determination: an update*. *PLoS One*, **4**, e5094.

Wiltzius, J. J., Sievers, S. A., Sawaya, M. R. & Eisenberg, D. (2009). *Atomic structures of IAPP (amylin) fusions suggest a mechanism for fibrillation and the role of insulin in the process*. *Protein Sci.* **18**, 1521–1530.

Yanez, M. E., Korotkov, K. V., Abendroth, J. & Hol, W. G. (2008). *The crystal structure of a binary complex of two pseudopilins: EpsI and EpsJ from the type 2 secretion system of Vibrio vulnificus*. *J. Mol. Biol.* **375**, 471–486.

Yau, S.-T., Petsev, D. N., Thomas, B. R. & Vekilov, P. G. (2000). *Molecular-level thermodynamic and kinetic parameters for the self-assembly of apoferritin molecules into crystals*. *J. Mol. Biol.* **303**, 667–678.

Ye, J.-D., Tereshko, V., Frederiksen, J. K., Koide, A., Fellouse, F. A., Sidhu, S. S., Koide, S., Kossiakoff, A. A. & Piccirilli, J. A. (2008). *Synthetic antibodies for specific recognition and crystallization of structured RNA*. *Proc. Natl Acad. Sci. USA*, **105**, 82–87.

Yip, C. K., Kimbrough, T. G., Felise, H. B., Vuckovic, M., Thomas, N. A., Pfuetzner, R. A., Frey, E. A., Finlay, B. B., Miller, S. I. & Strynadka, N. C. (2005). *Structural characterization of the molecular platform for type III secretion system assembly*. *Nature (London)*, **435**, 702–707.

Zhang, F., Basinski, M. B., Beals, J. M., Briggs, S. L., Churgay, L. M., Clawson, D. K., DiMarchi, R. D., Furman, T. C., Hale, J. E., Hsiung, H. M., Schoner, B. E., Smith, D. P., Zhang, X. Y., Wery, J.-P. & Schevitz, R. W. (1997). *Crystal structure of the obese protein leptin-E100*. *Nature (London)*, **387**, 206–209.

International Tables for Crystallography (2012). Vol. F, Chapter 4.4, pp. 140–144.

Chapter 4.4. High-throughput X-ray crystallography

K. H. Choi

4.4.1. Introduction

Structural-genomics projects have contributed to major developments in automation, miniaturization and process integration in X-ray crystallography. In the structural-genomics approach, multiple open reading frames from a genome are separately cloned and expressed without prior knowledge of the structure or function of the encoded proteins. Essential steps in a typical project include: (1) bioinformatics analysis of genome sequences for potential targets, (2) gene amplification by PCR (polymerase chain reaction) and subcloning into an appropriate expression vector, (3) protein expression and purification, (4) protein crystallization, (5) X-ray data collection, and (6) data processing and structure determination. Structural-genomics projects typically operate under a 'lowest hanging fruit' philosophy, pursuing structural targets that prove to be the most amenable to crystallization. However, for individual investigators working on specific problems, it is often necessary to focus on a particular protein. Many of the protocols developed in high-throughput (HT) structural genomics can also be adapted for parallel production of multiple constructs of a single protein target which is difficult to crystallize. This chapter summarizes recent developments in HT X-ray crystallography approaches and their application to parallel production of a single protein (Gräslund, Nordlund *et al.*, 2008; Manjasetty *et al.*, 2008; Sharff & Jhoti, 2003).

4.4.2. Design of multiple constructs: bioinformatics analysis of genome sequences

Often, the two primary bottlenecks in crystal structure determination are obtaining a soluble protein and the initial crystallization. If initial crystallization trials do not lead to successful crystallization, diffraction-quality crystals or structure determination, modification of the protein itself is often more successful in producing useful crystals than exhaustive screening/optimization of protein crystallization conditions. In the HT approach, generation of stable domains having variable N or C termini can be performed in parallel, vastly reducing time and resource requirements. The clones can be screened systematically for soluble, active proteins amenable to crystallization. This approach is particularly suited to problems in which one has no estimate, *a priori*, of which constructs are likely to be most successful at the expression, purification or crystallization steps.

It may be difficult to predict which modifications will influence protein solubility and the protein's ability to crystallize. A common and often useful strategy is to crystallize a similar protein from an alternative biological source. If an alternative source is either unavailable or equally problematic, gene fusion with large affinity tags can be used for increasing protein expression and solubility. Several fusion proteins, including thioredoxin, maltose binding protein, glutathione S-transferase, intein, SUMO (small ubiquitin-like modifier) protein and calmodulin-binding protein, have been used to generate soluble proteins. However, the fused tag may prevent crystallization because of conformational heterogeneity resulting from a flexible linker; hence, the tag must often be removed in an additional purification step. Protein engineering by amino-acid substitutions and chemical modifications (such as methylation of exposed lysines) have been shown to improve crystallization of some proteins (Rayment, 1997; Walter *et al.*, 2006; Wingren *et al.*, 2003). Large, flexible solvent-exposed residues (*e.g.* Lys or Glu) on the surface of proteins are substituted with smaller residues (*e.g.* Ala) to facilitate the formation of intermolecular contacts which stabilize the crystal lattice (the surface-entropy reduction approach; Cooper *et al.*, 2007; Derewenda, 2004). Random mutagenesis following selection of a desired phonotype (*e.g.* folding ability using a GFP (green fluorescent protein) reporter or solubility assays) has been used to produce soluble proteins from insoluble wild-type proteins (the directed-evolution approach; Pédelacq *et al.*, 2002). Since symmetric molecules such as homodimers may crystallize more readily than monomeric protein, new Cys residues can be introduced in a monomeric protein to induce homodimer formation *via* intermolecular disulfide bond formation (the synthetic symmetrization approach; Banatao *et al.*, 2006).

Bioinformatics tools, including multiple sequence alignment and sequence motif searches, as well as prediction of secondary structures, domain boundaries, membrane spanning and disordered regions, can be used to aid the rational design of constructs. General considerations in designing truncation mutations are to avoid truncations in the middle of predicted secondary structural elements, to avoid hydrophobic residues at the termini, and to eliminate membrane-spanning regions in the construct design. In the case of multi-domain proteins, truncation of the protein to smaller functional domains can be effective. Exact locations of the beginning and ending of a domain are still difficult to predict even with domain search programs, and thus biochemical approaches such as limited proteolysis can aid the determination of the domain boundaries. The optimal step size for truncation mutations cannot be predicted, but protein constructs varying by approximately five residues in length often show significant differences in solubility and crystallization behaviour (Choi *et al.*, 2004; Gräslund, Sagemark *et al.*, 2008).

4.4.3. Cloning

Development of robotics that utilize a 96-well format has changed traditional sequential cloning of individual proteins such that parallel HT cloning and protein preparation are now possible. The conventional steps in cloning are PCR amplification, restriction enzyme digestion, ligation, transformation, selection of transformers and protein expression tests. Many procedures, including amplification of target genes, cloning and screening for expression, can be achieved in parallel (96 samples at a time) using either a single liquid-handling robot or multichannel pipettors. In 96-well parallel cloning, reaction steps for individual wells cannot be optimized, and thus care should be taken to synchronize all reactions and to minimize the number of steps to increase efficiency. For this reason, certain cloning strategies are more popular for HT cloning. Recombinant protein

Copyright © 2012 International Union of Crystallography

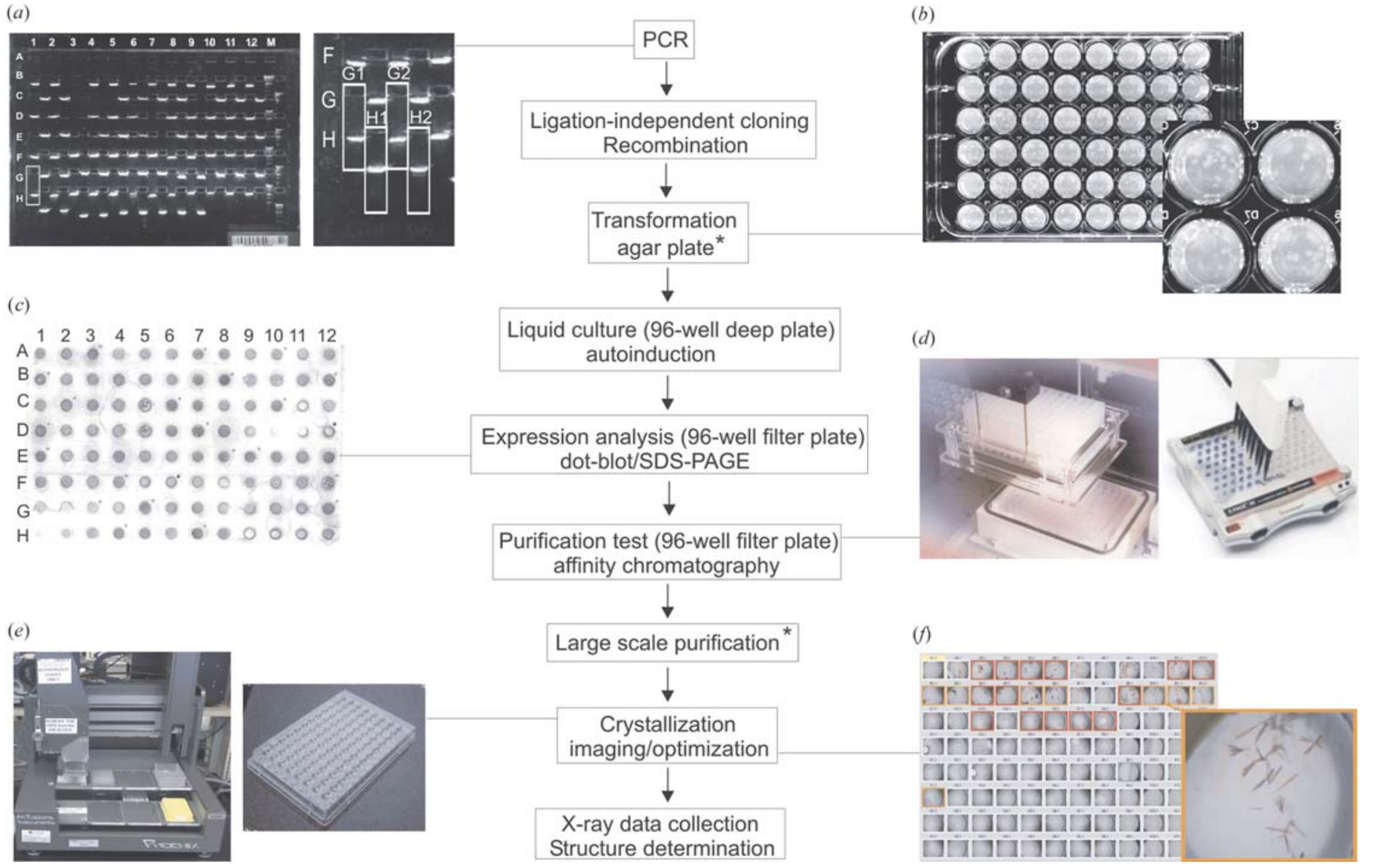

(a)
1 2 3 4 5 6 7 8 9 10 11 12 M
A B C D E F G H
F G H
G1 G2 H1 H2
PCR
Ligation-independent cloning
Recombination
Transformation
agar plate*
Liquid culture (96-well deep plate)
autoinduction
Expression analysis (96-well filter plate)
dot-blot/SDS-PAGE
Purification test (96-well filter plate)
affinity chromatography
Large scale purification*
Crystallization
imaging/optimization
X-ray data collection
Structure determination
(b)
(c)
1 2 3 4 5 6 7 8 9 10 11 12
A B C D E F G H
(d)
(e)
(f)

Figure 4.4.3.1
Diagram of HT cloning, expression and purification. All steps can be performed in a 96-well format except the step labelled with an asterisk (*). (*a*) PCR amplification. PCR products are loaded onto an E-gel (1% agarose) in a 96-well format with a molecular marker (lane 'M') and visualized with ethidium bromide. A close-up view of the gel outlining individual wells is also shown. (*b*) Transformation. Transformed cells were plated in 48-well agar plates. (*c*) Protein expression was tested using a western dot blot. H1 and H7 are negative and positive controls, respectively. (*d*) Small-scale purification using affinity resin. A 96-well filter plate in small-scale protein purification steps (left) and a 96-well pre-cast protein gel electrophoresis system (right) are shown. (*e*) Crystallization robot for setting up 96-well crystallization plates (right). (*f*) Crystal imaging and scoring system. A close-up view of an image of well B10 is shown on the right.

expression in *E. coli* is the most common, and will be discussed here.

4.4.3.1. Ligation-independent and recombination cloning

Ligation-independent cloning (LIC) strategies remove the need for restriction enzyme digestion and ligation of PCR products, and are thus ideal for use in an HT cloning procedure. In LIC, PCR primers are designed to append sequences that, after treatment with T4 DNA polymerase, generate 12- to 15-base overhangs that are complementary to overhangs in the vector (Aslanidis & de Jong, 1990). These longer cohesive ends make the insert–vector complex sufficiently stable to allow the transformation of hosts without ligation of the fragments.

The recombination strategy is based on the site-specific recombination reaction involved in bacteriophage λ integration and excision (Hartley *et al.*, 2000). PCR primers are designed to contain the specific sequences of recombination at $5'$ and $3'$ ends of the target gene, and the resulting PCR product is subcloned into a shuttle vector *via* site-specific recombination in the presence of integrase, integration host factor proteins and excisionase (*e.g.* the Gateway cloning system from Invitrogen). This clone can then either be isolated after transformation or directly used without purification for cloning into an expression vector. The major advantage of recombination cloning is that it provides a convenient way to shuttle an insert from one vector to another and thus is useful to test multiple expression conditions.

4.4.3.2. Practical application

Since the PCR reaction must be synchronized for all 96 wells in a plate, PCR primer design is important for the success of the PCR reaction. All PCR primers should have similar melting temperatures, between 50 and 60 °C. PCR products and the prepared vector are mixed and transferred into competent cells aliquoted into a 96-well plate. The PCR products and other DNA samples during the cloning, including the restriction-enzyme digest or a plasmid preparation (if needed), can be analysed *via* gel electrophoresis using a commercially available 96-well agarose gel (E-gel 96 from Invitrogen) in less than 15 min. The gel is compatible with a 96-well format and can be loaded either with a liquid-handling robot or multi-channel pipettor (Fig. 4.4.3.1*a*). The PCR step usually results in success rates as high as 98%. DNA purification kits are available in a 96-well format that use a vacuum manifold. As described above, the restriction-enzyme digest and ligation steps are not necessary in some cloning strategies.

Following transformation, cells are then plated into two 48-grid agar growth plates mixed with appropriate antibiotics (Fig. 4.4.3.1*b*). The 48-grid agar plate is made by inserting a cloning grill into a square Petri dish before pouring the agar, and the grill remains embedded in the media. Small glass beads can be added to each segment of the 48 grids to spread the solution. Colonies are picked by a colony picker, and are grown in 96-well deep-well blocks containing media and appropriate antibiotics. Using a liquid-handling robot and a colony picker for a single round of cloning, it is possible to achieve >80% efficiency in a 96-well format. The same robotic setup can be used for the subsequent recovery of recombinant plasmid before DNA sequencing and protein expression.

4.4.4. Protein expression and purification

Small-scale protein expression in an HT format enables evaluation of the expression, solubility and purification of target proteins or multiple constructs. Autoinduction allows many protein expression tests in parallel without having to monitor cell densities to optimize induction, and is thus ideal for HT approaches. Each protein in a 96-well plate will have different chemical properties and sizes, and thus affinity tags are the method of choice for parallel purification of multiple proteins on a single platform. Dot blots can be used to test expression or solubility, and miniaturized resin deposited in a 96-well plate can be used for a purification test. Although up to 90 µg of protein have been purified from 1 ml of *E. coli* cell culture (Scheich *et al.*, 2003), the amount of protein obtained from small-scale purification is generally not enough for initial crystallization trials, even with low-volume crystallization, and thus scaled-up protein production and purification is needed. The purification tags can be removed upon large-scale purification.

4.4.4.1. Autoinduction

Protein expression under control of the T7 *lac* promoter system can be induced either with the chemical inducer isopropyl-β-galactoside (IPTG) or by autoinduction using a mixture of glucose, glycerol and lactose during *E. coli* growth. In autoinduction, cells grow to relatively high density in a defined ratio of glucose-to-lactose media (Studier, 2005). Initially, glucose prevents induction by lactose. When available glucose is used up, lactose then induces target protein expression switching to lactose metabolism.

Cell cultures (1 ml) are grown in 2.2 ml deep-well 96-well blocks. Protein expression is autoinduced using commercially available glucose/lactose media (*e.g.* Overnight Express from EMD Biosciences). The following day, cells are lysed either by repeated freezing and thawing cycles, or by addition of a lysozyme solution. A combination of lysozyme and benzonase solutions (*e.g.* PopCulture reagent from Novagen) eliminates the need for cell harvesting prior to lysis and lysate clarification following lysis. The cell lysate can then be directly used for expression and purification tests. Alternatively, whole-cell lysates can be filtered though a 96-well filter plate, allowing for separation of inclusion bodies from the soluble fraction (filtrate) so that protein expression in the soluble fraction can be assessed.

4.4.4.2. Expression and solubility test: dot blot

Dot blot is a simple method that can be used to analyse either total protein expression in the cell lysate, or soluble protein expression following the separation of supernatant and cell pellet. Protein samples from the total cell lysate in a 96-well plate are dotted onto a nitrocellulose membrane by applying a vacuum. Target proteins are then probed with an antibody against the protein or against an affinity tag, *e.g.* anti-His-tag antibody for His-tagged protein detection (Fig. 4.4.3.1*c*). Alternatively, the cell lysates are analysed by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) for protein expression using a pre-cast 48- or 96-well SDS gel (E-PAGE 48 or 96 from Invitrogen). The gel consists of 48 or 96 wells for samples, and an additional four or eight wells for protein markers. The protein can be loaded either with a liquid-handling robot or with a multichannel pipettor, and electrophoresis is completed within 15 min (Fig. 4.4.3.1*d*).

4.4.4.3. Small-scale purification

Affinity purifications are preferred in a 96-well format because the specific interaction between the protein and affinity resin allows for a simple ‘bind–wash–elution’ procedure. Purification using a His-tag or a GST-tag (where GST = glutathione S-transferase) is popular in structure determination projects, and appropriate resins are available as magnetic beads or agarose resin in a 96-well plate, *i.e.* magnetic Ni-NTA or GST beads, or Ni- or Co-linked agarose discs, respectively.

The soluble fraction or whole-cell lysate is transferred to a 96-well plate containing an affinity resin for protein binding. The beads are separated from unbound protein by placing them in a magnetic stand designed to accommodate a 96-well format, or by filtration using a 96-well filter plate that retains the beads but allows passage of the cell lysate. Beads are washed, and bound protein is then eluted with the appropriate elution buffer (*i.e.* imidazole and reduced glutathione for His-tagged and GST-tagged proteins, respectively) into a 96-well plate by applying a vacuum (either as part of an appropriately equipped liquid-handling robot, or by using a 96-well vacuum manifold) or by centrifugation. Purified proteins are analysed by SDS–PAGE using a pre-cast 48- or 96-well protein gel. The gels can then be stained either by protein staining or western blotting. A new system for protein and DNA analysis has been developed based on a microfluidic chip (Caliper Labchip, Caliper Life Science). These chips contain a network of miniaturized channels, through which fluids and chemicals are moved to separate DNAs or proteins. The DNA and protein signal is measured by laser-induced fluorescence. The instrument is capable of separating 2 ng–2 µg amounts of protein, one well after another; it takes about 1 h to analyse a 96-well plate.

4.4.5. Crystallization

HT crystallization using 96-well crystallization plates greatly reduces the total amount of protein required for screening and thus structure determination. Robotic crystallization systems are capable of dispensing nanolitre droplets (<100 nl) and hence substantially increase the number of conditions that can be screened with a fixed amount of protein sample, as well as reducing the time required for setting up a series of crystallization trials. A thousand conditions with different crystallization parameters (*e.g.* pH, salts, temperature) can be screened with 100 µl of protein sample. At the time of writing, commercially available crystallization robots include the Honey Bee (Cartesian Micro-array), Phoenix (Rigaku), Mosquito (Molecular Dimensions Limited) and Oryx (Douglas Instrumentation). These crystallization robots can set up either vapour-diffusion, microbatch or hanging-drop methods in 96-well plates within 15 min. Crystallization robots can also be integrated into a larger system that includes storage and imaging of crystallization plates, and liquid-formulation robots.

Honey Bee and Phoenix robots contain a single (or several) non-contact channel that dispenses protein solution and a 96-channel dispenser head that dispenses the crystallization solutions (50 nl to 100 µl). The 96-channel head transfers crystallization solutions from a 96-well deep-well plate into the reservoir and crystallization drop in a 96-well crystallization plate (Fig. 4.4.3.1*e*). A single channel transfers protein solution into each of the 96-well drops one by one without touching the precipitate drops. The plate is then sealed with a clear film by the user. The Mosquito liquid-handling robot can set up drops with a hanging-drop geometry, and is more popular with membrane protein crystallization (detergent solutions have a tendency to adhere to the side of the drop well in sitting-drop geometry). The Mosquito uses only disposable tips capable of dispensing 20 nl–1.2 µl volume; thus the user can control the location of drop deposition more precisely because the disposable tips can touch the drop. The use of disposable tips also prevents cross-contamination between samples, and washing steps between samples are eliminated.

The Fluidigm TOPAZ system utilizes a new technology, the crystallization screen chip, in which protein sample and reagent solutions are automatically loaded into diffusion chambers within the protein screen chip and the two solutions mixed by free interface diffusion, as opposed to vapour diffusion or microbatch techniques (Thorsen *et al.*, 2002). A very small amount of protein is required for a crystallization screen, *i.e.* as little as 1.0 µl protein solution for 96 trials. Crystals obtained from the protein chip are generally too small for X-ray data collection, and thus need to be scaled up to obtain diffraction-quality crystals.

Progress of crystallization trials from large numbers of 96-well plates can be monitored using an imaging robot to take pictures of individual crystallization drops. The resulting images can then be analysed either manually or using automatic crystal recognition systems at specified time intervals (Markley *et al.*, 2009). Remote viewing of recorded crystal pictures is also available over the web. Each recorded image is linked to crystallization conditions for evaluation and scoring of the crystallization conditions (Fig. 4.4.3.1*f*). Minstrel (Rigaku), CrystalFarm (Bruker) or HomeBase (The Automation Partnership) systems offer integrated systems for plate storage and imaging.

Crystallization conditions that initially produced crystals should be optimized to improve crystal growth and quality. A liquid-handling robot can be used to make screens in 96-well deep-well plates. A liquid-formulation robot has been developed for protein crystallization to make grid screens of 96-well deep-well conditions (*e.g.* Alchemist from Rigaku). The crystallization conditions stored in an imaging robot are linked to the liquid-formulation software, and can be used to formulate 96-well screen conditions for optimization experiments. Other optimization methods such as crystallization in gels, control of nucleation using oil mixtures or microporous materials, and seeding experiments can also be employed in an HT fashion (Chayen, 2003; Georgiev *et al.*, 2006; Sugahara *et al.*, 2008).

4.4.6. Synchrotron data collection

High-brilliance beamlines at modern synchrotrons have significantly reduced the time required for X-ray data collection, and complete data sets can often be collected within minutes. Thus, the time required for crystal mounting and centring is no longer negligible. Automatic crystal mounting and centring allow users to remotely mount crystals for crystal evaluation and data collection without entering the experimental hutch (Manjasetty *et al.*, 2008; Sharff & Jhoti, 2003; Sugahara *et al.*, 2008).

Automated crystal mounting allows the screening of many crystals for diffraction quality and then goes back to the best diffraction-quality crystals for full data collection. The mounting robot picks up frozen crystals in a pin from a Dewar, puts them on a goniometer and retrieves the pin after the diffraction test. The automated crystal-mounting robots have primarily been developed for use at synchrotron sources, although a commercial version of an automatic sample-mounting robot is now available that can be used with a home-source X-ray generator (ACTOR from Rigaku and cryogenic sample changer from Marresearch). Fully automated crystal alignment is not yet available, but semi-automated crystal centring by clicking a mouse to indicate the intended centre of the crystal is used at most beamlines.

References

Aslanidis, C. & de Jong, P. J. (1990). *Ligation-independent cloning of PCR products (LIC-PCR). Nucleic Acids Res.* **18**, 6069–6074.

Banatao, D. R., Cascio, D., Crowley, C. S., Fleissner, M. R., Tienson, H. L. & Yeates, T. O. (2006). *An approach to crystallizing proteins by synthetic symmetrization. Proc. Natl Acad. Sci. USA*, **103**, 16230–16235.

Chayen, N. E. (2003). *Protein crystallization for genomics: throughput versus output. J. Struct. Funct. Genomics*, **4**, 115–120.

Choi, K. H., Groarke, J. M., Young, D. C., Rossmann, M. G., Pevear, D. C., Kuhn, R. J. & Smith, J. L. (2004). *Design, expression, and purification of a flaviviridae polymerase using a high-throughput approach to facilitate crystal structure determination. Protein Sci.* **13**, 2685–2692.

Cooper, D. R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M. & Derewenda, Z. (2007). *Protein crystallization by surface entropy reduction: optimization of the SER strategy. Acta Cryst.* D**63**, 636–645.

Derewenda, Z. S. (2004). *Rational protein crystallization by mutational surface engineering. Structure*, **12**, 529–535.

Georgiev, A., Vorobiev, S., Edstrom, W., Song, T., Laine, A., Hunt, J. & Allen, P. (2006). *Automated streak-seeding with micromachined silicon tools. Acta Cryst.* D**62**, 1039–1045.

Gräslund, S., Nordlund, P. *et al.* (2008). *Protein production and purification. Nat. Methods*, **5**, 135–146.

Gräslund, S., Sagemark, J., Berglund, H., Dahlgren, L. G., Flores, A., Hammarström, M., Johansson, I., Kotenyova, T., Nilsson, M., Nordlund, P. & Weigelt, J. (2008). *The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. Protein Expr. Purif.* **58**, 210–221.

Hartley, J. L., Temple, G. F. & Brasch, M. A. (2000). *DNA cloning using in vitro site-specific recombination. Genome Res.* **10**, 1788–1795.

Manjasetty, B. A., Turnbull, A. P., Panjikar, S., Bussow, K. & Chance, M. R. (2008). *Automated technologies and novel techniques to accelerate protein crystallography for structural genomics. Proteomics*, **8**, 612–625.

Markley, J. L., Aceti, D. J., Bingman, C. A., Fox, B. G., Frederick, R. O., Makino, S., Nichols, K. W., Phillips, G. N., Primm, J. G., Sahu, S. C., Vojtik, F. C., Volkman, B. F., Wrobel, R. L. & Zolnai, Z. (2009). *The Center for Eukaryotic Structural Genomics. J. Struct. Funct. Genomics*, **10**, 165–179.

Pédelacq, J. D., Piltch, E., Liong, E. C., Berendzen, J., Kim, C. Y., Rho, B. S., Park, M. S., Terwilliger, T. C. & Waldo, G. S. (2002). *Engineering soluble proteins for structural genomics. Nat. Biotechnol.* **20**, 927–932.

Rayment, I. (1997). *Reductive alkylation of lysine residues to alter crystallization properties of proteins. Methods Enzymol.* **276**, 171–179.

Scheich, C., Sievert, V. & Büssow, K. (2003). *An automated method for high-throughput protein purification applied to a comparison of His-tag and GST-tag affinity chromatography. BMC Biotechnol.* **3**:12

Sharff, A. & Jhoti, H. (2003). *High-throughput crystallography to enhance drug discovery. Curr. Opin. Chem. Biol.* **7**, 340–345.

Studier, F. W. (2005). *Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif.* **41**, 207–234.

Sugahara, M., Asada, Y., Shimizu, K., Yamamoto, H., Lokanath, N. K., Mizutani, H., Bagautdinov, B., Matsuura, Y., Taketa, M., Kageyama, Y., Ono, N., Morikawa, Y., Tanaka, Y., Shimada, H., Nakamoto, T., Yamamoto, M. & Kunishima, N. (2008). *High-throughput crystallization-to-structure pipeline at Riken SPring-8 center. J. Struct. Funct. Genomics*, **9**, 21–28.

Thorsen, T., Maerkl, S. J. & Quake, S. R. (2002). *Microfluidic large-scale integration. Science*, **298**, 580–584.

Walter, T. S., Meier, C., Assenberg, R., Au, K. F., Ren, J., Verma, A., Nettleship, J. E., Owens, R. J., Stuart, D. I. & Grimes, J. M. (2006). *Lysine methylation as a routine rescue strategy for protein crystallization. Structure*, **14**, 1617–1622.

Wingren, C., Edmundson, A. B. & Borrebaeck, C. A. (2003). *Designing proteins to crystallize through beta-strand pairing. Protein Eng.* **16**, 255–264.

International Tables for Crystallography (2012). Vol. F, Chapter 5.1, pp. 145–151.

PART 5. CRYSTAL PROPERTIES AND HANDLING

Chapter 5.1. Crystal morphology, optical properties of crystals and crystal mounting

H. L. Carrell and J. P. Glusker

5.1.1. Crystal morphology and optical properties

When crystals of a biological macromolecule are grown, they are first examined under the microscope. This can show the crystal quality and may reveal crystal symmetry. Some external properties of macromolecular crystals are described here, including information on shape, habit, polymorphism, twinning and the indexing of crystal faces. Optical properties are also described. Every crystal form, however, has to be treated individually; only by working with it can the crystallographer discover its physical properties. The diffractionist aims to be able to mount a macromolecular crystal in a totally stable manner so that it does not deteriorate or slip in position during the data collection. Some methods for mounting such macromolecular crystals for X-ray diffraction studies are described here together with the necessary tools. For general information on purchasing supplies to do this see the list at http://www.hamptonresearch.com.

5.1.1.1. Crystal growth habits

5.1.1.1.1. The shape of a crystal – growth habits

Morphology is the general study of the overall shape of a crystal, that is, the arrangement of faces of a crystal. It can often provide useful information about the internal symmetry of the arrangement of atoms within the crystal (Mighell *et al.*, 1993). The periodicity of the arrangement of molecules or ions in a crystal can be represented by three non-collinear vectors, **a**, **b** and **c**, which give a unit cell in the form of a parallelepiped with axial edges a, b and c, and interaxial angles α, β and γ (α between **b** and **c**, *etc.*). The vectors **a**, **b** and **c** from the chosen origin of the unit cell are, by convention, selected in a right-handed system. Since there may be several possible choices of unit cell, the simplest, with the smallest possible repeats and with interaxial angles nearest to 90°, is the best choice. One method used to highlight the periodicity of the atomic arrangement within a crystal is to replace each unit cell by a point; this mathematical construction gives the crystal lattice. The entire crystal structure is the convolution of the unit-cell contents with the crystal lattice.

Biological macromolecules are, in general, chiral and can only crystallize in those space groups that do not contain symmetry operations that would convert a left-handed molecule into a right-handed one (improper symmetry operations). Proper symmetry operations involve translations, rotation axes and screw axes. These maintain the chirality of the molecule and hence are appropriate for crystals of biological macromolecules. The number of possible space groups is therefore reduced by this constraint on the types of symmetry operations allowed from the usual 230 for molecules in general down to 65 for chiral molecules.

The appearance of a crystal that has grown under a particular set of experimental conditions is called its habit. It is a result of the different relative growth rates of various crystal faces, and these rates, in turn, depend on the nature of the interactions between the molecules in the crystal, the degree of supersaturation of the solution and the presence of any impurities which may affect the growth rates of specific crystal faces. The term 'habit' is only used to describe the various appearances of crystals that are composed of identical material and maintain the same unit-cell dimensions and space group. The faces that have developed on these crystals are various subsets of those implied in the overall morphological description of the crystal. Any change in the experimental conditions under which a crystal is grown may alter its habit; a judicious selection of experimental conditions may permit formation of crystals with a chunky habit that are more suitable for X-ray diffraction analysis than thin plates or needle-like crystals. Examples of the crystalline forms of haemoglobins are provided by Reichert & Brown (1909).

Various descriptions of crystal habits appear in the literature. These include terms such as 'tabular', 'platy' or 'acicular' crystals, 'hexagonal rods' and 'truncated tetragonal bipyramids', among others. Some crystal habits are not very appropriate for X-ray diffraction analyses; these include spherulites, which are polycrystalline aggregates of fine needles with an approximately radial symmetry, and dendrites, which have a tree-like structure. The habit of a crystal can sometimes give information on the molecular arrangement within it. For example, flat molecules that stack readily upon each other produce long crystalline needles, because interactions in the stacking direction are stronger than those in other directions.

A crystal is bounded by those faces that have grown most slowly. Fast-growing faces quickly disappear as more and more molecules are deposited on them, constrained by surrounding faces that are growing more slowly. Any factor that changes the relative rates of growth of crystal faces, such as impurities in the crystallizing solution, will affect the overall habit. Different faces of protein crystals have different arrangements of side chains on their surfaces; thus, an impurity may bind to certain faces rather than others. Adsorption of an impurity on a particular face of a crystal may retard the growth of that face, causing it to become more prominent than normal in the growing crystal.

5.1.1.1.2. Quality of protein crystals

Protein and nucleic acid crystals contain a high proportion of water in each unit cell and are therefore fragile. The proportion of solvent to macromolecule in the crystal can be expressed, as described by Matthews (1968), as V_m in $\text{\AA}^3\,\text{Da}^{-1}$ for the asymmetric unit. Values in the range 1.7 to 4.0 are usual for proteins, but nucleic acid crystals generally have a higher water content. Crystal fragility due to water content may be used to determine whether or not a crystal contains protein or buffer salt. Pressure with a fine probe will settle this question because a protein crystal will shatter, while a salt crystal, which is much sturdier, will generally withstand such treatment. If crystals have grown into one another, or appear as clumps, it is sometimes possible to split off a single crystal by prodding the clump gently at the junction point between the crystals with a scalpel or a glass fibre.

Copyright © 2012 International Union of Crystallography

5.1.1.1.3. *Polymorphism*

Intermolecular contacts between protein molecules in the crystalline state determine the mechanical stability of the crystal. If the conditions used for crystallization vary, the number and identity of these contacts may be changed, and polymorphs will result. Polymorphism is the existence of two or more crystalline forms of a given material. Polymorphs have different unit-cell dimensions and hence different molecular arrangements within them. This property is common for biological macromolecules and can be used to select the best crystalline form for X-ray diffraction studies. Different polymorphs of a particular material are often prepared by varying the crystallization conditions. They may also develop in the same crystallizing drop because supersaturation conditions may change while a crystal is growing. Examples of polymorphs are provided by hen egg-white lysozyme, which can form tetragonal, triclinic, monoclinic or orthorhombic crystals, depending on the pH, temperature and nature of added salts in the crystallization setup (Steinrauf, 1959; Ducruix & Giegé, 1992: Oki *et al.*, 1999).

A regular surface that offers a charge distribution pattern that is complementary to a possible protein layer in a crystal can sometimes be useful in producing a starting point for the nucleation of new protein crystals. Epitaxy is the oriented growth of one material on a crystal of an entirely different material. Regularities on the surface of the first crystal can act as a nucleus for the oriented growth of the second material. Generally, there should be similar, but not necessarily exactly matching, repeat distances in the two crystals. Epitaxy has been used with considerable success for the growth of protein crystals on selected mineral surfaces (McPherson & Shlichta, 1988). For example, lysozyme crystals grow well on the surface of the mineral apophyllite. Crystals of related macromolecules can also be used as nucleation sources for protein crystallization. Epitaxy can, however, sometimes be a nuisance rather than a benefit if the crystallization setup contains surfaces with unwanted regularity.

A change in the environment around a protein crystal may also cause a change in unit-cell dimensions, and possibly even in space group. For example, the transference of a RuBisCO crystal from a high-salt, low-pH mother liquor to a low-salt, high-pH synthetic mother liquor produced a more densely packed polymorph. The overall unit-cell dimensions were smaller in the latter (V_m changed from 3.16 to 2.74 $\text{Å}^3\ \text{Da}^{-1}$) (Zhang & Eisenberg, 1994).

5.1.1.1.4. *Twinning*

Twinning is a phenomenon that can cause much grief in X-ray diffraction data measurements. It has been described as 'a crystal growth anomaly in which the specimen is composed of separate crystal domains whose orientations differ in a specific way ... some or all of the lattice directions in the separate domains are parallel' (Yeates, 1997). Thus, a twin consists of two (or more) distinct but coalescent crystals. This effect has been described in terms of their diffraction patterns as follows: 'Some crystals show splitting of diffraction spots owing to the different tilts of the two lattices. Others pretend to be single crystals with no split spots, and their symmetries of intensity distribution vary with every data set. These latter have been called hemihedral, and in them the unique axes of the two crystals are exactly reversed parallel with each other.' (Igarashi *et al.*, 1997). Perfect hemihedral twinning (when there are equal proportions of each twin member) can be detected from the value of $\langle I^2\rangle/\langle I\rangle^2$ for the acentric data; it is near 2 for untwinned data and 1.5 for twinned data (Yeates, 1997).

Twinning may sometimes be prevented by introducing variations into the crystallization setup. Such changes could involve the pH or the nature of the buffer. Variations in the seeding technique used, or the introduction of additional agents, such as metal ions or salts, detergents, or certain amino acids, can also be tried. One method for estimating the degree of twinning is based on the fact that each measured X-ray diffraction intensity is the sum of the intensities from the two (or more) crystal lattices (suitably weighted according to the proportion in which each lattice alignment occurs in the crystal). The relative proportion of each component can therefore be estimated. Detwinned intensities obtained by this method should only be positive or zero (not negative) within experimental error (Stanley, 1972; Britton, 1972; Rees, 1980). Crystal structures of hemihedrally twinned crystals are now being determined (Gomis-Rüth *et al.*, 1995; Breyer *et al.*, 1999).

5.1.1.2. *Properties of crystal faces*

5.1.1.2.1. *Indexing crystal faces*

Crystal faces are described by three numbers, the Miller indices, that define the relative positions at which the three axes of the unit cell are intercepted by the crystal face (Blundell & Johnson, 1976; Phillips, 1957). The crystal face that is designated *hkl* makes intercepts $x = a/h$, $y = b/k$ and $z = c/l$ with the three axes (a, b and c) of the unit cell. When the Miller indices are negative, because the crystal face intercepts an axis in a negative direction, they are designated $-h - k - l$ or $\bar{h}\bar{k}\bar{l}$. Most faces of a crystal are, as required by the law of rational indices, represented by small integers. For symmetry reasons, planes in hexagonal crystals are conveniently described by four axes, three in a plane at 120° to each other. This leads to four indices, *hkil*, where $i = -(h + k)$. Crystal faces are designated by round brackets, *e.g.* (100), as shown in two simple examples in Fig. 5.1.1.1. A set of faces then defines a crystal form. All sets of planes *hkl* related by symmetry, such as the main faces of an octahedron (111), $(11\bar{1})$, $(1\bar{1}1)$, $(\bar{1}11)$, $(1\bar{1}\bar{1})$, $(\bar{1}1\bar{1})$, $(\bar{1}\bar{1}1)$ and $(\bar{1}\bar{1}\bar{1})$, may be represented by the use of curly brackets, *e.g.* {111}. Square brackets, *e.g.* [111], are used to indicate a direction in a crystal, [001] being the *c* axial direction. Thus, information about several faces of a crystal is contained in information about one face if the relevant symmetry of the crystal is known. If the atomic structure of the crystal is known, it is possible to determine which aspects of the macromolecule lie on the various crystal faces.

5.1.1.2.2. *Measurement of crystal habit*

The habit of a crystal may be described in detail by measurements, particularly of the angles between adjacent faces. This can be done with an optical goniometer, in which a collimated visible beam of light is reflected from different faces as the crystal is rotated, and the angles between positions of high light intensity are noted. Such studies can be combined with X-ray diffraction measurements of the crystal orientation of the unit cell with respect to the various crystal faces (Oki *et al.*, 1999).

5.1.1.3. *Optical properties*

Crystals interact with light in a manner which depends on the arrangement of atoms in the crystal structure, any symmetry in this arrangement and the chemical nature of the atoms involved. Refraction is seen as a change in the direction of a beam of light when it passes from one medium to another (such as the apparent bend in a pencil placed in a beaker of water). The refractive index is the ratio of the velocity of light in a vacuum to that in the material under investigation and is greater than unity. It is measured by the extent to which the direction of a beam of light changes on entering a medium.

If a protein crystal has grown in the cubic system, its refractive index will be the same in all directions and the crystal is described as optically isotropic. Most protein crystals, however, form crystals with anisotropic properties, that is, their properties vary with the direction of measurement of the crystal. For example, some crystals appear differently coloured when viewed in different directions, and are described as pleochroic. The absorption of light is greatest when light is vibrating along bonds of chromophoric groups in the molecules in the crystals rather than perpendicular to them, and therefore they show interesting effects in plane-polarized light (in which the electric vectors lie in one plane only) (Bunn, 1945; Wahlstrom, 1979).

Anisotropy of the refractive index of a protein crystal, that is, its birefringence, can be used by biochemists to determine whether or not a protein has been crystallized. If the protein preparation in a test tube is held up to the light and shaken, birefringent protein microcrystals are revealed by light streaks, a schlieren effect. This occurs because the crystals have a different refractive index from the bulk of the liquid. Birefringence implies that there is double refraction as light passes through a crystal, and the light is split into two components (the ordinary and extraordinary rays) that travel with different velocities and have different properties (those of the ordinary ray being normal). Iceland spar (calcite) provides the ideal example of double refraction. Birefringence is measured as the difference between the refractive indices for the ordinary and extraordinary rays, and a crystal is described as positively birefringent if the refractive index is greater for the extraordinary ray. If a crystal is positively birefringent ($n_E > n_O$), it can be assumed to contain rod-like bodies lying parallel to the single vibration direction of greatest refractive index (n_E). If a crystal is negatively birefringent ($n_O > n_E$), it can be assumed to contain plate-like bodies lying perpendicular to the single vibration direction of least refractive index (n_E). For example, in crystalline naphthalene, the highest refractive index is in the direction of the highest density of atoms, along the long axis of the molecule. Similar arguments can be applied to crystalline macromolecules, such as haemoglobin, which is strongly pleochroic, appearing dark red and opaque in two extinction directions, and light red and transparent in the third (Perutz, 1939). Thus, the coefficient of absorption is high when the electric vibration of plane-polarized light is parallel to the haem groups, but is low in other directions. Similarly, a specific carotenoid protein has been found to appear orange or clear depending on the orientation of the crystal relative to the direction of polarization of the light hitting it (Kerfeld *et al.*, 1997). The darkest orange colour, corresponding to a maximum absorbance of the carotenoid cofactor, is found when the polarizer is aligned along the *a* axis (the long axis of the crystals). This suggests that all the carotenoid cofactor molecules in this crystal structure lie nearly parallel to the *a* axis.

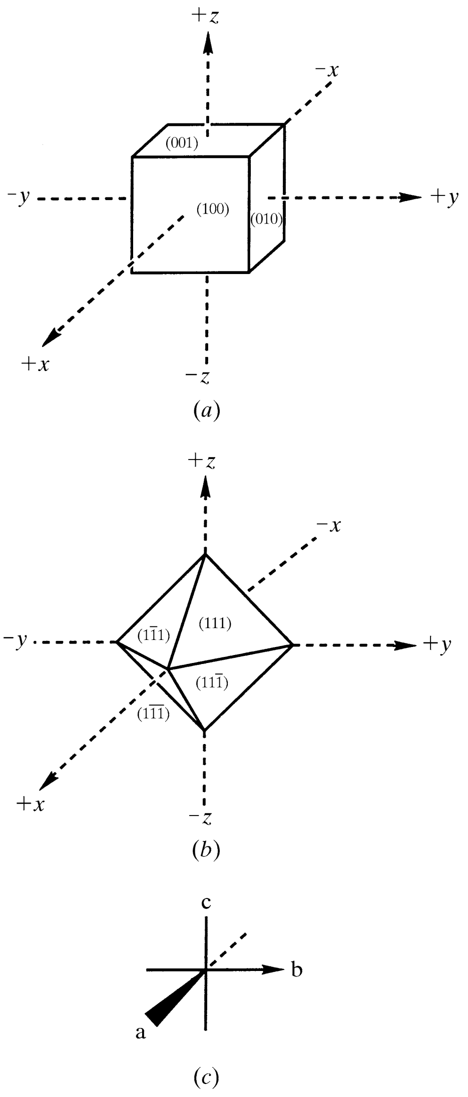


Figure 5.1.1.1
Crystal faces. (*a*) Cube and (*b*) octahedron. (*c*) Unit-cell axes.

The directions along which double refraction is observed can be used to give some information on the crystal class. Some crystals are found to have one, and only one, direction (the optic axis) along which there is no double refraction. Crystals with this property are called uniaxial. They have two principal refractive axes and are tetragonal, hexagonal or rhombohedral. Other crystals are found to have two directions along which there is no double refraction (two optic axes), and these are called biaxial. Such crystals are either orthorhombic, monoclinic or triclinic and have three principal refractive indices.

5.1.1.3.1. Crystals between crossed polarizers

Most protein crystals are birefringent and are brightly coloured in polarized light. In order to view these effects, crystals (in their mother liquor) are set on a microscope stage with a Nicol prism (polarizing material) between the light source and the microscope slide (the polarizer). Another Nicol prism is set between the crystal and the eyepiece (the analyser). The crystal should not be in a plastic container, since this would produce too many colours. If the vibration plane of the analyser is set perpendicular to that of the polarizer (to give 'crossed Nicols'), no light will pass through in the absence of crystals, and the background will be dark. If the crystal is isotropic, the image will remain dark as the crossed Nicol prisms are rotated. If, however, the crystal is birefringent (with two refractive indices), the crystal will appear coloured except at four rotation positions (90° apart) of the crossed Nicol prisms, where the crystal and background will be dark (extinguished). At these positions, the vibration directions of the Nicol prisms coincide with those of the crystal. If one is looking exactly down a symmetry axis of a crystal that is centrosymmetric in projection (such as a tetragonal or hexagonal crystal), the crystal will not appear birefringent, but dark. By noting the external morphology of the crystal with respect to its angle of rotation, one can often deduce the directions of the unit-cell axes in the crystal (Hartshorne & Stuart, 1960). Examination of a crystal under crossed Nicol prisms can also provide information on crystal quality. For example, sometimes the components of a twinned crystal extinguish plane-polarized light independently. Other methods of examining crystals include Raman spectroscopy (Kudryavtsev *et al.*, 1998).

5.1.1.3.2. Refractive indices and what they tell us about structure

The refractive index of a crystal can be measured by immersing it in a mixture of liquids of a known refractive index in which the crystal is insoluble. The liquid composition is then varied until the crystal appears invisible. At this point, the refractive indices of the crystal and the liquid are the same. If the refractive index is the same in all directions, the crystal is optically isotropic, but most protein crystals are optically anisotropic and have more than one refractive index. For example, tetragonal crystals have different refractive indices for light vibrating parallel to the fourfold axis and for light vibrating perpendicular to it. These refractive indices are measured by the use of plane-polarized light.

5.1.1.4. Packing of molecules in crystals

Growth kinetics of the different faces should be correlated with the structural anisotropy of the intermolecular contacts. It has been found that a judicious mutation of a single surface residue of a protein can markedly affect its solubility and hence crystallizability. This method has been used with great success for crystallizing a retroviral integrase (Dyda *et al.*, 1994).

The relationship between crystal morphology and internal crystal structure was examined in the mid-1950s (Hartman & Perdok, 1955*a*,*b*,*c*). It was shown that the morphology of a crystal is determined by 'chains' of strong intermolecular interactions (hydrogen bonding, van der Waals contacts, molecular stacking) running through the entire crystal. For a crystal to grow in the direction of a strong interaction ('bond'), these bonds must form an uninterrupted chain through the structure, giving rise to the periodic bond chain theory. The stronger the interaction between molecules, the more likely the crystal is to be elongated in that direction. If a bond chain contains interactions of different kinds, its influence on the shape of the crystal is determined by the weakest interaction present in a particular chain. Prominent faces are parallel to at least two high-energy bond chains. This enables a correlation to be made between the crystal lattice and the crystal morphology, based on the fact that direct protein–protein contacts, reinforced by well ordered solvent molecules, are important in determining crystal packing (Frey *et al.*, 1988). Studies of the morphology of tetragonal lysozyme (Nadarajah & Pusey, 1996; Nadarajah *et al.*, 1997) showed that the crystallizing unit is a helical tetramer (centred on the 4_3 crystallographic axes).

5.1.2. Crystal mounting

5.1.2.1. Introduction to crystal mounting

Once crystals have been obtained and visually characterized, the next procedure involves the transfer of a selected crystal to an appropriate mounting device so that the crystal may be characterized using X-rays. Macromolecular crystals are generally obtained from and stored in a solution containing the precipitant or precipitants and other substances such as uncrystallized protein or other macromolecules. The object is to mount the crystal in such a way that it is undamaged by cracking, drying out, dissolving *etc.* during this operation. In some cases, the crystal may have been stored in a solution containing volatile solvents. Alternatively, the crystals may have been grown at a temperature lower than room temperature and therefore may require special handling in order to avoid crystal deterioration. In other cases, it may be desirable to prepare the crystal for study at cryogenic temperatures. This section deals with the mounting of crystals for all these conditions and concentrates on the mounting of crystals for diffraction experiments at or just below room temperature. Procedures such as 'flash cooling' are used to reduce radiation damage. Crystal-mounting techniques for cryogenic experiments are covered in detail in Part 10 and are only mentioned briefly here. In general, the most difficult part of mounting macromolecular crystals is the transfer of the crystal from a holding solution to a suitable mount. A capillary or, if cryogenic experiments are to be carried out, a cryoloop should be used.

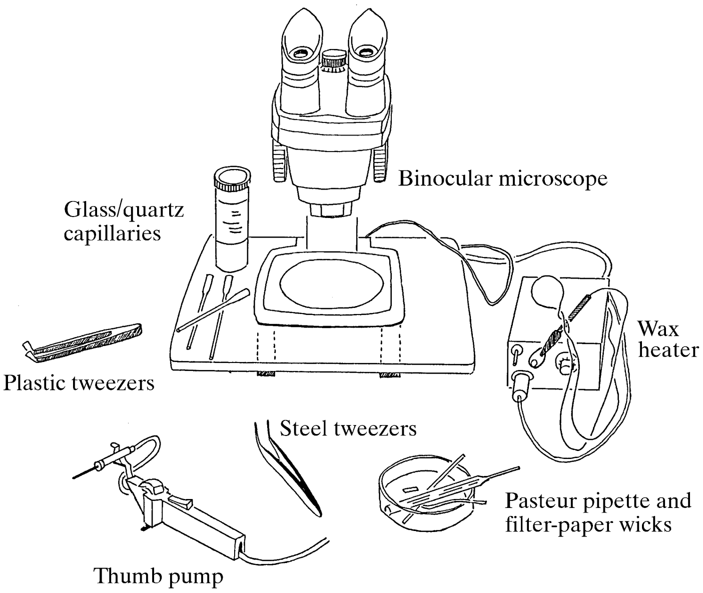


Figure 5.1.2.1
Tools commonly used for mounting crystals.

5.1.2.2. Tools for crystal mounting

In order to facilitate the process of mounting macromolecular crystals for X-ray diffraction experiments, it is necessary to have the appropriate tools for the task. Fig. 5.1.2.1 shows a collection of some useful tools for the mounting of crystals. These include a binocular microscope, tweezers (two types), thin glass capillaries, Pasteur pipettes, a heater, paper wicks, and a thumb pump. Other useful tools and supplies include surgical scissors, dental wax, latex tubing, light vacuum oil, a cryogenic mounting loop, Plasticine, mounting platforms, mounting pins, absorbent dental points and micropipettes with plastic tips. There are many other items that might be useful, and several variations are found in different laboratories. An important factor in the transfer of crystals from a holding solution to a capillary is that the experimenter needs to feel at ease with the process. The method that will be detailed here has evolved over time and has proved to be a relatively anxiety-free process. Other methods for crystal mounting may be found in the literature (Rayment, 1985; Sawyer & Turner, 1992; McRee, 1993). All of the methods outlined here and in the literature have the same goal, namely, the successful transfer of a macromolecular single crystal to a suitable mount for X-ray data collection.

5.1.2.2.1. *Microscope*

Perhaps the single most important piece of equipment for examining and mounting crystals is a binocular dissection microscope. This should have variable zoom capabilities, and there should be sufficient distance (*e.g.* 5–10 cm) between the objective lens of the microscope and the microscope stage to accommodate the necessary equipment and allow manipulation of the crystals and solutions. A magnification of between 10 and 40 times is probably best in practice. It is also important to ensure that the light source of the microscope is not so intense that it heats the microscope stage, thereby damaging the macromolecular crystals. If the microscope is fitted with crossed polarizers, the quality of the crystals can be assessed.

5.1.2.2.2. *Capillaries*

The capillaries used for crystal mounting are made of thin-walled glass. These capillaries range from 0.1 to 2.0 mm in diameter and have a stated wall thickness of 0.01 mm. In practice, however, the larger the diameter of the capillary, the thinner the glass wall. Therefore, handling of the larger-diameter capillaries is generally very difficult because they are so fragile. Capillaries made of fused quartz are also available, but are not recommended for general use because they produce a higher background with X-rays. Quartz capillaries are not as fragile as the thin-walled glass capillaries, however, and may be useful in experiments where the tensile strength of the capillary is important, for example, when a diffraction flow-cell experiment is planned (Petsko, 1985). In addition, small-diameter capillaries (produced in the laboratory by drawing out glass tubing or Pasteur pipettes) will be needed to aid in the removal of excess liquid around the crystal after the transfer from the crystallization dish to a capillary has been completed.

5.1.2.2.3. *Thumb pump*

The thumb pump is a simple micropipetting device for transferring very small amounts of liquid in a highly controlled manner, making it an extremely useful tool for directly transferring protein crystals from solution to capillary, thus minimizing the chance of crystal damage. This simple device allows the experimenter to have much more control over volume transfers than any other device we have tried. The mechanism is simple and easy to operate. The device shown in the illustration can be held and manipulated with one hand. The capillary is held firmly in place and the pipetting action is controlled by a thumb wheel (part of the thumb pump), which affords a great deal of control over the volume of liquid being transferred with the crystal.

5.1.2.2.4. *Heater*

The heater illustrated here consists of a variable rheostat and a heating element. The latter is a short piece of Nichrome wire which has been coiled and attached to the rheostat *via* wires that run through a ball-point pen barrel. This permits a fine temperature control for melting dental wax and for controlling where the heat is applied.

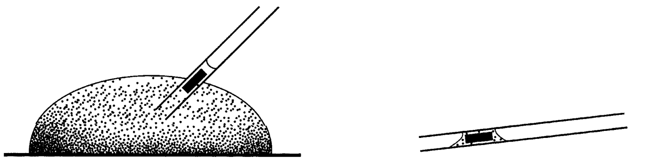
5.1.2.3. *Capillary mounting*

One must first select a capillary for mounting the crystal. A general rule to follow is to select a capillary that has a diameter that is approximately twice the size of the crystal dimension to be placed along the breadth of the capillary. Thus, to mount an elongated parallelepiped with the longest crystal dimension parallel to the capillary, it is necessary to take into account the cross section of the crystal perpendicular to the longest dimension. This rule is only a guide and is probably broken most of the time. Indeed, for ‘chunky’ crystals, it may be advantageous to use a capillary only slightly larger than the crystal so that the crystal may be in contact with the capillary wall in more than one place, thereby making the mount more stable. The object is to have enough of the crystal in contact with the capillary wall to allow the crystal to be held in place with a small amount of mother liquor. One possible problem that occurs with very thin crystals is that the crystal may bend to conform to the shape of the capillary wall. In this case, the crystal is rendered unsuitable for X-ray diffraction experiments.

The capillary is prepared by first removing the flame-sealed tip. This is done with surgical scissors or by pinching with surgical tweezers and then gently tapping the capillary tip against a smooth hard surface to remove the jagged edges which may have resulted from this cutting. The removal of the jagged edges at the broken end of the capillary will simplify the transfer of the crystal from the holding solution to the capillary. The large flared end of the capillary is left intact, and, at this time, a rubber transfer tube with a mouthpiece can be fitted over the flared end of the capillary. The capillary can then be rinsed with distilled deionized water, or it may even be desirable to treat the capillary with some other solution, such as EDTA, or perhaps with a solution identical to that surrounding the crystal. The rinse solution is gently drawn up into the capillary, and the solution is then blown out into a waste container. This can be accomplished fairly rapidly, and then the capillary can be dried by gently drawing air in through the capillary tip. Only the excess liquid should be removed at this point if the rinse solution contains salts.

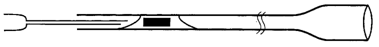
The crystal must now be transferred to a capillary from the storage location, which may be a shallow well in a depression slide, a droplet on a cover slip, or perhaps a vial containing crystals. Direct transfer from a droplet on a cover slip or from a shallow well of a depression slide to the final capillary is possible, but can be complicated if several crystals are present in the drop. The easiest way to set up the final crystal transfer is to first remove it from the original drop or vial using a micropipette with a tip that has been enlarged so that it will accommodate the desired crystal. The crystal is then transferred together with a few microlitres of solution to a siliconized cover slip or depression well using the micropipette. It may even be easier to place 5–10 µl of solution on a siliconized cover slip or depression well and then use a cryoloop to capture the crystal and deposit it in the solution. The crystal can now be easily drawn up into the capillary with the aid of the thumb pump. It will be accompanied by a small column of mother liquor, and the thumb pump can be used to position the crystal with its liquid at the desired location in the capillary. The excess mother liquor can now be removed by using a capillary that is much smaller than the data-collection capillary, as well as smaller than the crystal. A final drying can be accomplished using appropriately sized filter-paper wicks or absorbent dental points. A very small amount of liquid should be left behind to keep the crystal moist and to ‘glue’ the crystal to the capillary wall *via* surface tension. A crystal that is too dry will probably deteriorate and be useless for diffraction experiments, while a crystal that has too much liquid can slip during data collection. On the other hand, moderate drying has been found, in certain cases, to give a crystal with improved diffraction.

After the crystal is safely in position in the capillary, the capillary must be sealed in order to maintain the moisture necessary to prevent crystal deterioration. If desired, a short column of mother liquor may be placed in the capillary a few

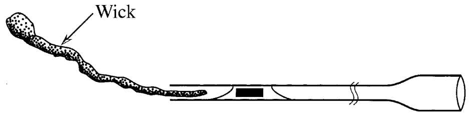


(*a*) Using the thumbwheel, draw the crystal into the capillary along with a short column of liquid

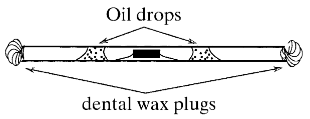
(*b*) Still using the thumbwheel, gently draw the crystal and mother liquor to the desired position in the capillary



(*c*) Using a small capillary, remove the excess mother liquor



(*d*) Remove liquid to desired dryness with a filter-paper wick



(*e*) Place columns of oil 4–5 mm from each end of the crystal and seal the ends of the capillary with dental wax

Figure 5.1.2.2
Mounting a crystal in a capillary.

millimetres away from the crystal. This is usually necessary if capillaries larger than 1 mm in diameter are used. A small strip of filter paper may also be placed in the capillary and then dampened with mother liquor. Both methods allow the moisture level in the crystal to be maintained. A reasonably good first seal may consist of a short column of light vacuum oil on both sides of the crystal, again, a few millimetres away from it. At this time, a ring of molten dental wax is placed along the capillary beyond the oil drop nearest the flared end of the capillary, and the capillary is then cut or broken just beyond the wax. The final seal may then be accomplished using molten dental wax or perhaps even epoxy at each end of the capillary. The diffraction equipment and arrangement will dictate the position of the crystal in the capillary, and this should be accommodated before the final seals are put in place. The geometry of the capillary could aid in preventing slippage of the wedged crystal during data collection (Åkervall & Strandberg, 1971). Alternatively, a specific crystal coating which effectively glues the crystal to the interior of the capillary can be used (Rayment *et al.*, 1977). The capillary with its crystal is now ready to be placed on the platform of choice for placement on the goniometer head in final preparation for diffraction experiments. Fig. 5.1.2.2 illustrates the steps in mounting a crystal in a capillary in preparation for the X-ray experiment.

The above method deals only with crystals which are to be mounted at or near room temperature for experiments at or near room temperature. An alternative approach is to grow a crystal in a capillary (Åkervall & Strandberg, 1971), which could eliminate the need to manipulate the crystal manually. When crystals have been grown in the presence of detergents or gels, specific methods may be required for mounting (McRee, 1993). The appropriate procedure for flash cooling of crystals is detailed in Part 10.

The authors are supported by grant CA-10925 from the National Institutes of Health.

References

Åkervall, K. & Strandberg, B. (1971). *X-ray diffraction studies of the satellite tobacco necrosis virus. III. A new crystal mounting method allowing photographic recording of 3 Å diffraction data. J. Mol. Biol.* **62**, 625–627.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York, London, San Francisco: Academic Press.

Breyer, W. A., Kingston, R. L., Anderson, B. F. & Baker, E. N. (1999). *On the molecular-replacement problem in the presence of merohedral twinning: structure of the N-terminal half-molecule of human lactoferrin. Acta Cryst.* D**55**, 129–138.

Britton, D. (1972). *Estimation of twinning parameter for twins with exactly superimposed reciprocal lattices. Acta Cryst.* A**28**, 296–297.

Bunn, C. W. (1945). *Chemical Crystallography. An Introduction to Optical and X-ray Methods*. Oxford: Clarendon Press.

Ducruix, A. & Giegé, R. (1992). Editors. *Crystallization of Nucleic Acids and Proteins. A Practical Approach*. Oxford, New York, Tokyo: IRL Press.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R. & Davies, D. R. (1994). *Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science*, **266**, 1981–1986.

Frey, M., Genovesio-Taverne, J.-C. & Fontecilla-Camps, J. C. (1988). *Application of the periodic bond chain (PBC) theory to the analysis of the molecular packing in protein crystals. J. Cryst. Growth*, **90**, 245–258.

Gomis-Rüth, F. X., Fita, I., Kiefersauer, R., Huber, R., Avilés, F. X. & Navaza, J. (1995). *Determination of hemihedral twinning and initial structural analysis of crystals of the procarboxypeptidase A ternary complex. Acta Cryst.* D**51**, 819–823.

Hartman, P. & Perdok, W. G. (1955*a*). *On the relations between structure and morphology of crystals. I. Acta Cryst.* **8**, 49–52.

Hartman, P. & Perdok, W. G. (1955*b*). *On the relations between structure and morphology of crystals. II. Acta Cryst.* **8**, 521–524.

Hartman, P. & Perdok, W. G. (1955*c*). *On the relations between structure and morphology of crystals. III. Acta Cryst.* **8**, 525–529.

Hartshorne, N. H. & Stuart, A. (1960). *Crystals and the Polarising Microscope. A Handbook for Chemists and Others*, 3rd ed. London: Edward Arnold & Co.

Igarashi, N., Moriyama, H., Mikami, T. & Tanaka, N. (1997). *Detwinning of hemihedrally twinned crystals by the least-squares method and its application to a crystal of hydroxylamine oxidoreductase from Nitrosomonas europaea. J. Appl. Cryst.* **30**, 362–367.

Kerfeld, C. A., Wu, Y. P., Chan, C., Krogmann, D. W. & Yeates, T. O. (1997). *Crystals of the carotenoid protein from Arthrospira maxima containing uniformly oriented pigment molecules. Acta Cryst.* D**53**, 720–723.

Kudryavtsev, A. B., Mirov, S. B., DeLucas, L. J., Nicolete, C., van der Woerd, M., Bray, T. L. & Basiev, T. T. (1998). *Polarized Raman spectroscopic studies of tetragonal lysozyme single crystals. Acta Cryst.* D**54**, 1216–1229.

McPherson, A. & Shlichta, P. (1988). *Heterogeneous and epitaxial nucleation of protein crystals on mineral surfaces. Science*, **239**, 385–387.

McRee, D. E. (1993). *Practical Protein Crystallography*, pp. 21–28. San Diego, New York: Academic Press.

Matthews, B. W. (1968). *Solvent content in protein crystals. J. Mol. Biol.* **33**, 491–497.

Mighell, A. D., Rodgers, J. R. & Karen, V. L. (1993). *Protein symmetry: metric and crystal (a precautionary note). J. Appl. Cryst.* **26**, 68–70.

Nadarajah, A., Li, M. & Pusey, M. L. (1997). *Growth mechanism of the (110) face of tetragonal lysozyme crystals. Acta Cryst.* D**53**, 524–534.

Nadarajah, A. & Pusey, M. L. (1996). *Growth mechanism and morphology of tetragonal lysozyme crystals. Acta Cryst.* D**52**, 983–996.

Oki, H., Matsuura, Y., Komatsu, H. & Chernov, A. A. (1999). *Refined structure of orthorhombic lysozyme crystallized at high temperature: correlation between morphology and intermolecular contacts. Acta Cryst.* D**55**, 114–121.

Perutz, M. F. (1939). *Absorption spectra of single crystals of haemoglobin in polarized light. Nature (London)*, **143**, 731–733.

Petsko, G. (1985). *Flow cell construction and use. Methods Enzymol.* **114**, 141–146.

Phillips, F. C. (1957). *An Introduction to Crystallography*, 2nd ed. London, New York, Toronto: Longmans Green & Co.

Rayment, I. (1985). *Treatment and manipulation of crystals*. *Methods Enzymol.* **114**, 136–140.

Rayment, I., Johnson, J. E. & Suck, D. (1977). *A method for preventing crystal slippage in macromolecular crystallography*. *J. Appl. Cryst.* **10**, 365.

Rees, D. C. (1980). *The influence of twinning by merohedry on intensity statistics*. *Acta Cryst.* A**36**, 578–581.

Reichert, E. T. & Brown, A. P. (1909). *The Differentiation and Specificity of Corresponding Proteins and Other Vital Substances in Relation to Biological Classification and Organic Evolution: the Crystallography of Hemoglobins.* Washington DC: Carnegie Institution of Washington. (Publication No. 116.)

Sawyer, L. & Turner, M. A. (1992). *X-ray analysis*. In *Crystallization of Nucleic Acids and Proteins*. *A Practical Approach*, edited by A. Ducruix & R. Giegé, pp. 267–274. Oxford, New York, Tokyo: IRL Press.

Stanley, E. (1972). *The identification of twins from intensity statistics*. *J. Appl. Cryst.* **5**, 191–194.

Steinrauf, L. K. (1959). *Preliminary X-ray data for some new crystalline forms of β-lactoglobulin and hen egg-white lysozyme*. *Acta Cryst.* **12**, 77–79.

Wahlstrom, E. E. (1979). *Optical Crystallography*, 5th ed. New York, Chichester, Brisbane, Toronto: John Wiley.

Yeates, T. O. (1997). *Detecting and overcoming crystal twinning*. *Methods Enzymol.* **276**, 344–358.

Zhang, K. Y. J. & Eisenberg, D. (1994). *Solid-state phase transition in the crystal structure of ribulose 1,5-bisphosphate carboxylase/oxygenase*. *Acta Cryst.* D**50**, 258–262.

International Tables for Crystallography (2012). Vol. F, Chapter 5.2, pp. 152–157.

Chapter 5.2. Crystal-density measurements

E. M. Westbrook

5.2.1. Introduction

Crystal-density measurements have traditionally been a valuable and accurate (±4%) method for determining molecular weights of proteins (Crick, 1957; Coleman & Matthews, 1971; Matthews, 1985). But since exact chemical compositions of proteins being crystallized today are usually known from DNA sequences, crystal densities are rarely used for this purpose. Rather, crystal-density measurements may be necessary to define a crystal's molecular-packing arrangement, particularly when a crystal has an unusual packing density (very dense or very open); when there are a large number of subunits in the crystallographic asymmetric unit; when the structure consists of heterogeneous subunits, so the molecular symmetry or packing is uncertain; and for crystals of nucleic acids, nucleic acid/protein complexes and viruses.

5.2.2. Solvent in macromolecular crystals

Crystals of biological macromolecules differ from crystals of smaller molecules in that a significant fraction of their volume is occupied by solvent (Adair & Adair, 1936; Perutz, 1946; Crick, 1957). This solvent is not homogeneous: a part binds tightly to the macromolecule as a hydration shell, and the remainder remains free, indistinguishable from the solvent surrounding the crystal.

Hydration is essential for macromolecular stability: bound solvent is part of the complete macromolecule's structure (Tanford, 1961). Diffraction-based studies of macromolecular crystals verify the presence of well defined bound solvent. Typically, 8–10% of the atomic coordinates in each Protein Data Bank file are those of bound water molecules. The consensus observation of protein hydration (Adair & Adair, 1936; Perutz, 1946; Edsall, 1953; Coleman & Matthews, 1971; Kuntz & Kaufmann, 1974; Scanlon & Eisenberg, 1975) is that every gram of dry protein is hydrated by 0.2–0.3 g of water: this is consistent both with the presence of a shell of hydration, the thickness of which is about one water molecule (2.5–3 Å), and with the rule-of-thumb that approximately one water molecule is found for every amino-acid residue in the protein's crystal structure. Matthews (1974) suggests setting this hydration ratio, w, to 0.25 g water per gram of protein as a reasonable estimate for typical protein crystals.

Crystallographic structures also exhibit empty regions of 'free' solvent. Such voids are to be expected: closely packed spheres occlude just 74% of the space they occupy, so to the extent that proteins are spherical, tight packing in their crystals would leave 26% of the crystal volume for free solvent. Although the distinction between free and bound solvent is not sharp (solvent-binding-site occupancies vary, as do their refined B factors), it is a useful convention and is consistent with many observed physical properties of these crystals.

5.2.3. Matthews number

In an initial survey of 116 crystals of globular proteins (Matthews, 1968) and in a subsequent survey of 226 protein crystals (Matthews, 1977), Matthews observed that proteins typically occupy between 22 and 70% of their crystal volumes, with a mean value of 51%, although extreme cases exist, such as tropomyosin, whose crystals are 95% solvent (Phillips *et al.*, 1979). The volume fraction occupied by a macromolecule, φ_m, is reciprocally related to V_M, the Matthews number, according to

$$V_M = \overline{v}_m / N_o \varphi_m = V/nM, \tag{5.2.3.1}$$

where $\overline{v}_m$ is the partial specific volume of the macromolecule (Tanford, 1961), N_o is Avogadro's number, V is the volume of the crystal's unit cell, n is the number of copies of the molecule within the unit cell and M is the molar weight of the macromolecule (grams per mole). V_M is the ratio between the unit-cell volume and the molecular weight of protein contained in that cell. The distribution of V_M (2.4 ± 0.5 Å^3 Da^{-1}) is asymmetric, being sharply bounded at 1.7 Å^3 Da^{-1}, a density limit consistent with spherical close packing. The upper limits to V_M are much less distinct, particularly for larger proteins. Matthews observed a slight tendency for V_M to increase as the molecular weight of proteins increases. V_M values below 1.9, or above 2.9, can occur but are relatively rare (beyond a 1σ cutoff).

The unit-cell volume, V, is determined from crystal diffraction. The partial specific volume of a macromolecule, $\overline{v}_m$, is the rate of change in the volume of a solution as the (unhydrated) macromolecule is added. It can be measured in several ways, including by ultracentrifugation (Edelstein & Schachman, 1973) and by measuring the vibrational frequency of a capillary containing a solution of the macromolecule (Kratky *et al.*, 1973). $\overline{v}_m$ typically has a value around 0.74 cm^3 g^{-1} for proteins and around 0.50 cm^3 g^{-1} for nucleic acids (Cantor & Schimmel, 1980). Values of $\overline{v}_m$ are tabulated for all amino acids and nucleotides, and $\overline{v}_m$ of a macromolecule can be estimated with reasonable accuracy as the mean value of its monomers. Commercial density-measuring instruments are available to determine $\overline{v}_m$ by the Kratky method. Because M is usually well known from sequence studies, n – the number of copies of the macromolecule in the unit cell – can be calculated thus:

$$n = V/V_M M = V N_o \varphi_m / \overline{v}_m M. \tag{5.2.3.2}$$

For proteins, evaluating this expression with $V_M = 2.4$ usually provides an unambiguous integer value for n – which must be a multiple of the number of general positions in the crystal's space group! Setting n to its integer value then provides the actual value for V_M. If the calculated V_M value lies beyond the usual distribution limits, if n has an unexpected value or a large value, or if the crystal contains unusual components or several different kinds of molecular subunits, the crystal density may need to be measured accurately.

5.2.4. Algebraic concepts

Let V be the volume of one unit cell of the crystal. Let m_c be the total mass within one unit cell, and m_m, m_{bs} and m_{fs} be the masses, within one unit cell, of the macromolecule, bound solvent and free solvent, respectively. Let ρ_c, ρ_m, ρ_{bs} and ρ_{fs}, respectively, be the densities of a complete macromolecular crystal, its unsolvated macromolecule, its bound-solvent compartment and its free-solvent compartment. Let φ_m, φ_{bs} and φ_{fs}, respectively, be the

Copyright © 2012 International Union of Crystallography

fractions of the crystal volume occupied by the unsolvated macromolecule, the bound solvent and the free solvent. By conservation of mass,

$$m_c = m_m + m_{bs} + m_{fs}. \tag{5.2.4.1}$$

The volume fractions must all add to unity:

$$\varphi_m + \varphi_{bs} + \varphi_{fs} = 1. \tag{5.2.4.2}$$

The density of the crystal is the total mass divided by the unit-cell volume:

$$\rho_c = \frac{m_c}{V} = \frac{m_m}{V} + \frac{m_{bs}}{V} + \frac{m_{fs}}{V}. \tag{5.2.4.3}$$

The mass in each solvent compartment is the product of its density and the volume it occupies:

$$m_{bs} = \rho_{bs} V \varphi_{bs}, \qquad m_{fs} = \rho_{fs} V \varphi_{fs}. \tag{5.2.4.4}$$

The mass of the macromolecule in the cell can be defined either from its partial specific volume, $\overline{\upsilon}_m$, the unit-cell volume, V, and the molecules' volume fraction, φ_m, or from the molar weight, M, the number of molecular copies in the unit cell, n, and Avogadro's number, N_o:

$$m_m = V\varphi_m/\overline{\upsilon}_m = nM/N_o. \tag{5.2.4.5}$$

Now (5.2.4.3) may be rewritten as

$$\rho_c = \varphi_m/\overline{\upsilon}_m + \rho_{bs}\varphi_{bs} + \rho_{fs}\varphi_{fs}. \tag{5.2.4.6}$$

Define a mean solvent density, ρ_s:

$$\rho_s = \frac{\rho_{bs}\varphi_{bs} + \rho_{fs}\varphi_{fs}}{\varphi_{bs} + \varphi_{fs}}. \tag{5.2.4.7}$$

This allows (5.2.4.6) to be rewritten as

$$\rho_c = \varphi_m/\overline{\upsilon}_m + (1 - \varphi_m)\rho_s. \tag{5.2.4.8}$$

Upon rearrangement, this gives expressions for the volume fraction of a macromolecule and for the molecular-packing number:

$$\begin{aligned} \varphi_m &= \frac{\rho_c - \rho_s}{(\overline{\upsilon}_m)^{-1} - \rho_s} = \frac{n\overline{\upsilon}_m M}{VN_o}, \\ n &= \frac{VN_o}{\overline{\upsilon}_m M}\frac{\rho_c - \rho_s}{(\overline{\upsilon}_m)^{-1} - \rho_s}. \end{aligned} \tag{5.2.4.9}$$

In (5.2.4.9), all terms can be measured directly, except ρ_s. The treatment of ρ_s will be discussed in Section 5.2.7. (5.2.4.9) defines the total macromolecular mass in the unit cell, $m_m = nM/N_o$, from a measurement of the crystal density ρ_c. If M were known from the primary sequence of the molecule, this measurement determines the molecular-packing number, n, with considerable certainty. If the molar weight were not accurately known, it could be determined by measuring the crystal density.

5.2.5. Experimental estimation of hydration

During refinement of crystal structures, crystallographers must decide how many solvent molecules are actually bound to the macromolecule and for which refined coordinates are meaningful. The weight fraction of bound solvent to macromolecule in the crystal, w, is estimated for most protein crystals to be about 0.25 (Matthews, 1974, 1985). However, its true value can be derived experimentally in the following manner. Since all relevant studies identify the bound solvent as water, it is reasonable to set the density of bound solvent as $\rho_{bs} = 1.0$ g ml^{-1}. Therefore, w can be expressed algebraically as

$$w = \frac{m_{bs}}{m_m} = \frac{\varphi_{bs}\rho_{bs}\overline{\upsilon}_m}{\varphi_m} = \frac{\varphi_{bs}\overline{\upsilon}_m}{\varphi_m}. \tag{5.2.5.1}$$

For crystals in which the rules-of-thumb $w = 0.25$ and $\overline{\upsilon}_m = 0.74$ cm^3 g^{-1} are valid, (5.2.5.1) implies that bound solvent occupies about one-third of the volume occupied by protein.

The crystal density, ρ_c, changes linearly with the density of free solvent surrounding the crystal. Let ρ_o be defined as the density the crystal would have if all its solvent were pure water ($\rho_s = 1.0$ g ml^{-1}):

$$\rho_o = 1 + \varphi_m(1/\overline{\upsilon}_m - 1). \tag{5.2.5.2}$$

A plot of crystal density against density of the supernatant (free solvent) solution should be a straight line with an intercept (at $\rho_{fs} = 1.0$ g ml^{-1}) of ρ_o and a slope of φ_{fs}. Therefore, by making a few crystal-density measurements, each with the crystal first equilibrated in solutions of varying densities, experimental values for ρ_o and φ_{fs} can be derived. If the partial specific volume is known for this molecule, φ_m, φ_{bs} and w can be derived from the expressions above. This approach was used by Coleman & Matthews (1971) and Matthews (1974) to measure molecular weights of six crystalline proteins, assuming $w = 0.25$, but their measurements could alternatively have assigned more accurate values to w, had the molecular weights been previously known. Scanlon & Eisenberg (1975) measured w for four protein crystals by this method (values betwen 0.13 and 0.27 were observed) and also confirmed that bound solvent exhibited a density of 1.0 g ml^{-1}.

5.2.6. Methods for measuring crystal density

Density measurements of macromolecular crystals are complicated by their delicate constitution. These crystals tolerate neither dehydration nor thermal or physical shock or stress. Furthermore, since macromolecular crystals contain free solvent, their densities will change as the density of the solvent in which they are suspended is changed. They cannot be picked up with tweezers, nor rinsed with arbitrary solvents, nor placed out to dry on the table.

The other experimental problem with these crystals is that they are very small. Typically, their linear dimensions are 0.1–0.2 mm, volumes are 1–10 nl and weights are 1–10 µg. Molecular structures can now be determined from even smaller crystals (linear dimensions as small as 20 µm) using synchrotron radiation, so density-measurement methods compatible with very small crystals are required. With such small samples, it is far easier and more accurate to measure *densities* than to measure directly *volumes* and *weights*.

The physical properties of macromolecular crystals constrain the methods by which their densities can be measured accurately. In all circumstances, great care must be taken to avoid artifacts such as air bubbles or particulate matter which often adhere to these crystals. All measurements should be made at one tightly controlled temperature, since thermal expansion can change densities and thermal convection can corrupt density gradients. Because crystals contain solvent, it is bad to dry them, since this process usually disrupts them, changing all parameters in unpredictable ways. Yet many density-measurement methods require that all external solvent first be removed from the crystals, since the measured densities will be some average of crystal and any remaining solvent. This can be an almost insurmountable problem for crystals containing cavities and voids. Unfortunately, many crystallization mother liquors are viscous and difficult to remove, for example if they contain polyethylene glycol (PEG).

Richards & Lindley in Chapter 3.2 of *IT* C (2004) list six methods for measuring crystal densities: pycnometry, the method of Archimedes, volumenometry, the immersion microbalance, flotation and the gradient tube. The first three methods require direct weighing of the crystal and are therefore of limited value for crystals as small as those used in macromolecular diffraction, although these methods are used in various applications, such as mineralogy and the sugar industry. The latter three methods measure densities and density differences, and can therefore be used in macromolecular crystallography. A new method specifically for protein crystals has recently been described (Kiefersauer *et al.*, 1996), involving direct tomographic measurement of crystal volumes coupled with quantitative amino-acid analysis. Because the gradient-tube method remains the method of choice for most crystal-density measurements, it will be discussed last and most thoroughly here.

5.2.6.1. *Pycnometry*

Pycnometry measures the density of a liquid by weighing a calibrated volumetric flask before and after it is filled with the liquid. To measure a crystal's density, the pycnometer is first calibrated and weighed, and then the crystal sample, from which all external liquid has been removed, is introduced. The pycnometer is now reweighed, thus determining the crystal's weight. Next, liquid of known density is added, and the pycnometer is reweighed. The crystal volume is derived from the difference in volumes of the pycnometer with and without the crystal present. The method requires direct measurement of the crystal's weight, yet it is difficult to make microbalances with sensitivity and accuracy limits better than about 0.01 mg. Micropycnometry methods have been developed to determine mineral densities with as little as 5 mg of material (Syromyatnikov, 1935), but typical macromolecular crystals are 1000 times smaller than that.

5.2.6.2. *Volumenometry*

This technique measures the increase in gas pressure due to changes in the volume of a calibrated container into which the crystal, having previously been weighed, is introduced (Reilly & Rae, 1954). Almost any gas can be used for this technique if it is compatible with the apparatus and does not interact with the crystal. The empty, calibrated chamber is pressurized by adding a measured gas bolus, and this pressure is measured. After it is weighed, the crystal sample is placed in the container, which is repressurized by adding the same volume of gas. The difference in pressures between the two measurements is due to the change in volume of the container due to the crystal's presence. In principle, this method is compatible with powdered crystal samples, multiple crystals, or irregularly shaped crystals. As with micropycnometry, this technique is not appropriate for macromolecular crystals. The crystal must be free of external solvent, yet not dried, and the microbalance must be able to measure the crystal's weight precisely and accurately. The useful lower limit of crystal size for this technique, reported by Richards & Lindley in Chapter 3.2 of *IT* C (2004), is 0.01 ml.

5.2.6.3. *The method of Archimedes*

Known for thousands of years, this method measures the difference in weight of an object in air and in a liquid of known density. The difference divided by the liquid density gives the object's volume. The crystal is suspended by a vertical fibre or wire from a microbalance as it is dipped into the liquid. The surface tension of the liquid acting on the supporting fibre must be accounted for and corrected. The accuracy of this method improves as the density of the liquid used approaches that of the crystal. This method has been used with crystals as small as 25 mg (Berman, 1939; Graubner, 1986), but it does not lend itself well to density measurements of objects as small as macromolecular crystals.

5.2.6.4. *Immersion microbalance*

Barbara Low and Fred Richards developed this ingenious method, which permits the crystal to be weighed in a liquid environment. A microbalance (consisting of a thin horizontal quartz fibre, free at one end) is kept entirely within the liquid bath. Its vertical deflection, observed with a microscope, is initially calibrated as a function of weight (Low & Richards, 1952*b*; Richards, 1954). The density of the liquid can be determined with high precision by standard techniques. Each crystal's volume (in the 1952–1954 studies) was calculated from two orthogonal photomicrographs (this required that the crystal morphology be regular). The crystal density can then be calculated:

$$\rho_c = \rho_{\rm liquid} + \frac{\text{apparent crystal weight}}{\text{crystal volume}}. \qquad (5.2.6.1)$$

The easiest and most accurate part of the method is measuring the liquid density. Therefore, experimental error in determining crystal weight and volume can be minimized by using a liquid with a density close to that of the crystal. In the limit where the crystal and liquid densities are the same, this method is equivalent to the flotation method – the fibre deflection is zero and the accuracy of the crystal-density measurement should be high. As originally implemented, the method is useful only for crystals with simple shapes, for which orthogonal photomicrographs can yield good estimates for the volume. Perhaps the method might be generalized if the tomographic volume-measuring method were adopted, as described by Kiefersauer *et al.* (1996). Richards & Lindley in Chapter 3.2 of *IT* C (2004) state that the method is only suitable for large crystals (volumes of 0.1 mm^3 or greater).

5.2.6.5. *Flotation*

The crystal must first be wiped completely free of external liquid and then immersed in a mixture of organic solvents, the density of which is adjusted (by addition of denser or lighter solvents) until the crystal neither rises nor sinks. Note that if the liquid used were aqueous, the crystal density would change as the surrounding liquid density is changed (*e.g.* by adding salt), since the crystal's free-solvent compartment would exchange with the external liquid. In this case, the equilibrium density, ρ_e, is a function only of the hydration number, w, and the macromolecule's partial specific volume, $\overline{v}_m$:

$$\rho_e = (w + 1)/(w + \overline{v}_m). \qquad (5.2.6.2)$$

ρ_e is about 1.25 g ml^{-1} for all protein crystals, regardless of packing arrangements or molecular weights, since $w \simeq 0.25$ and $\overline{v}_m \simeq 0.74$ cm^3 g^{-1}.

When the crystal just floats, the liquid's density (which now equals the crystal density) can be measured by standard techniques with high accuracy. Flotation measurements can be made with small samples (Bernal & Crowfoot, 1934) and with slurries of microcrystals. Centrifugation should be used to accelerate the crystal settling rate each time the liquid density is altered. The

Table 5.2.6.1
Organic liquids for density determinations

Name	Density (g ml^{-1})
Carbon tetrachloride (tetrachloromethane), CCl_4	1.5940
Bromobenzene, CH_5Br	1.4950
Chloroform (trichloromethane), $CHCl_3$	1.4832
Methylene chloride (dichloromethane), CH_2Cl_2	1.3266
Chlorobenzene, CH_5Cl	1.1058
Benzene, CH_6	0.8765
m-Xylene, 1,3-$(CH_3)_2C_6H_4$	0.8642
Iso-octane (2-methylheptane), $C_5H_{11}CH(CH_3)_2$	0.6980

Table 5.2.6.2
Inorganic salts for density determinations

The densities are approximate values for aqueous solutions at 20 °C.

Solute	Density (g ml^{-1})
Sodium chloride	1.20
Potassium tartrate	1.40
Potassium iodide	1.63
Iron(III) sulfate	1.80
Zinc bromide	2.00
Zinc iodide	2.39

method can be tedious, so its practitioners rarely achieve an accuracy better than 0.2–1.0% (Low & Richards, 1952*a*).

5.2.6.6. *Tomographic crystal-volume measurement*

Recently, a new method for density measurement which is specific for protein crystals has been reported (Kiefersauer *et al.*, 1996). The crystal volume is calculated tomographically from a set of optical-shadow back projections of the crystal, with the crystal in many (>30) orientations. This measurement is analogous to methods used in electron microscopy (Russ, 1990). The crystal is mounted on a thin fibre which is in turn mounted on a goniostat capable of positioning it in many angular orientations. The crystal must remain bathed in a humidity-regulated air stream to avoid drying. The uncertainty of the volume measurement improves asymptotically as the number of orientations increases (estimated to be 10–15%). The images are captured by a digital charge-coupled device camera, transferred to a computer and processed with the program package *EM* (Hegerl & Altbauer, 1982). This same crystal must then be recovered and subjected to quantitative amino-acid analysis (the authors used a Beckman 6300 amino-acid analyser). With a lower limit of 100 pmol for each amino acid, the uncertainty of this measurement was estimated to be 10–20% for typical protein crystals. The method appears to work for crystals with volumes ranging between 4–50 nl. Errors in the determined values of n ranged from 4–30%.

Implementation of the method requires complex equipment and considerable commitment (in terms of hardware and software) by the research laboratory. The accuracy of the method is sufficient to determine n unambiguously in many cases, but it is not as high as can be obtained with gradient-tube or flotation methods if care is taken. The method has the virtue that once established in a research laboratory, it might lend itself to a considerable degree of automation, thereby reducing the activation barrier to measuring crystal densities for members of the research group.

5.2.6.7. *Gradient-tube method*

This is the most commonly used method for measuring densities of macromolecular crystals. It is simple and inexpensive to implement. It can be used to measure densities of very small crystals and crystalline powders. Practised with care, the gradient-tube method is capable of measuring crystal densities with a precision and accuracy of ±0.002 g ml^{-1}.

Although density gradients were used earlier for other purposes, the application of the gradient-tube method for crystal-density measurement was first described by Low & Richards (1952*a*). The gradient can be formed in a long glass column (preferably with volume markings, such as a graduated cylinder), in which case the crystal sample will settle by gravity in the tube; or in a transparent centrifuge tube, in which case the crystal's approach to its equilibrium density may be accelerated by centrifugation. The gradient may be made by two organic liquids (Table 5.2.6.1) with different densities, or it may be made by a salt concentration gradient in water (Table 5.2.6.2). In either case, formation of the gradient is simplified with a standard double-chamber 'gradient maker' – however, a *glass* gradient maker should be used if the gradient is made of organic solvents! Be aware that all these substances are toxic, particularly to the liver, and some are listed as carcinogens, so avoid prolonged exposure.

Desired upper and lower density limits for the gradient can be made by mixing two of these liquids in appropriate ratios. The sensitivity and resolution of the measurement can be enhanced by using a shallow gradient covering the expected density. These organic liquids have a nontrivial capacity to dessicate the crystal sample, so it is important that they be water-saturated before use. Also, when an alcohol is the precipitant of a crystal, organic solutions may be inappropriate for density measurements.

For aqueous gradients, the salts listed in Table 5.2.6.2 may be added to water to create a dense liquid.

A widely used variant of the method has been to form aqueous gradients with Ficoll, a sucrose polymer cross-linked with epichlorhydrin (Westbrook, 1976, 1985; Bode & Schirmer, 1985). Manufactured by the Pharmacia Corporation specifically for making density gradients used in the separation of intracellular organelles or intact cells, Ficoll is a large polymer ($M_r = 400\ 000$) which is very hydrophilic and soluble, and has chemical properties similar to sucrose. Since it is highly cross-linked, each Ficoll molecule tends to be globular and is so large that it is effectively excluded from the crystal. Ficoll precipitates protein from solution on a per-weight basis as effectively as polyethylene glycol and can prevent protein crystals from dissolving, even in the absence of other solutes. A 60% (w/w) solution of Ficoll has a density of about 1.26 g ml^{-1}, sufficiently dense that almost all protein crystals will float in this solution (nucleic acid crystals are usually too dense for Ficoll). Used with care (see below), Ficoll gradients seem to yield the most reproducible crystal-density measurements. Concentrated Ficoll solutions are quite viscous, so these gradients are usually made by manually overlaying small volumes (0.5 ml each) of decreasing density, rather than with a gradient maker. In a standard cellulose nitrate centrifuge tube of about 5 ml capacity, this procedure makes an almost continuous gradient which works satisfactorily.

The density column must be calibrated once it has been formed. This is performed by introducing small items of known density into the column and noting their vertical positions. The density of the gradient as a function of vertical position can then be defined by interpolating between adjacent calibrated points. Usually, the calibrating points are made from small drops of immiscible liquid. Thus, in an organic solvent gradient, the drops are made of salt water; in an aqueous gradient, the drops are made of mixed organics (previously saturated with water). To make each calibration drop, a solution is made up with approximately the desired density, and its exact density (± 0.002 g ml^{-1}) is measured pycnometrically or by refractive index (Midgley, 1951). The drops can be inserted into the gradient with a flame-narrowed Pasteur pipette (this takes practice). Once calibrated, these gradients tend to be extremely stable over many months.

With an organic liquid gradient, two methods have been used to introduce the crystal sample to be measured. It can be extracted from its mother liquor with a pipette and extruded onto filter paper, which wicks away all exterior aqueous liquid. When free of moisture, but before it dessicates, the crystal must be shaken, flipped, or scraped onto the gradient top surface and allowed to sink to its equilibrium position. The second method involves injection of the crystal sample, in an aqueous droplet, into the gradient solution with a Pasteur pipette. A very thin syringe (home-made or commercial) is then used to draw off all extraneous liquid, while the crystal remains submerged in the organic liquid. Either method requires considerable manual dexterity and practice, especially with very small crystals. A significant advantage of Ficoll and aqueous salt gradients is that the crystal does not need to be manipulated at all: any liquid surrounding the crystal, which was introduced into the gradient at the start, rapidly dilutes into the aqueous solution and does not appear to interfere with further measurements.

With very small crystals, the approach to equilibrium is so slow that it is wise to use centrifugation, especially if it is suspected that the density is changing with time (see below). Nitrocellulose centrifuge tubes compatible with swinging-bucket rotors are typically 1 cm diameter, 5 cm long cylinders and are suitably transparent for this work. Centrifugation at 2500–5000 r.p.m. for as little as five minutes is sufficient for most crystals to reach a stable position in the gradient. It can be difficult to find the crystal after centrifugation, so the one or two most likely density values should be calculated in advance, and looked for first. The positions of calibration drops and of crystals in these centrifuge tubes can be measured with a hand-held ruler to a resolution of about 0.5 mm.

Particularly for crystals with high values of V_M (*i.e.*, loosely packed) or for crystals of large molecular weight proteins, the apparent crystal density may increase with time: the crystal continues to sink and there is no apparent equilibrium spot. This behaviour is seen in both organic solvent gradients and Ficoll gradients, and the reasons for it are unclear. It may be that, in organic solvent gradients, some of the solvent can dissolve into the crystal; or the crystal may condense from slow dessication. In Ficoll gradients, it may be that sucrose monomers or dimers are present, which diffuse into the crystal over time. A careful study of this behaviour (Bode & Schirmer, 1985) in Ficoll gradients suggested that useful density values can still be obtained for these crystals by fitting the apparent density to an exponential curve:

$$\rho_c(t) = a + b\exp(-\lambda t). \qquad (5.2.6.3)$$

In this expression, parameters a, b and λ must be derived from the fitted curve. The crystals were inserted into the gradient with flame-narrowed Pasteur pipettes. Each crystal was initially surrounded by a small amount of mother liquor, which rapidly diffused into the Ficoll solution. Time zero was assigned as the time when centrifugation first began. It was necessary to observe crystal positions within the first minute, and at two- to five-minute intervals thereafter, to obtain a reasonable time curve for the density function. The experimental goal in the Bode & Schirmer experiment was to obtain a good estimate for the density value at time zero, $\rho_c(0) = a + b$. This was realized in all six of the crystal forms that manifested time-dependent density drift in the study.

5.2.7. How to handle the solvent density

It is necessary to have an accurate estimate of the mean solvent density, ρ_s, in (5.2.4.9). The Ficoll gradient-tube method is particularly convenient for this reason: the gradient can be made without any significant solute other than Ficoll. Since the free-solvent compartment of the crystal is entirely water, $\rho_s = \rho_{bs} = \rho_{fs} = 1.0$ g ml^{-1}. Therefore, in Ficoll density gradients, the crystal density becomes ρ_o, as defined in (5.2.5.2), and the packing number n can be calculated from

$$n = \frac{VN_o}{M}\frac{(\rho_c - 1)}{(1 - \overline{v}_m)}. \qquad (5.2.7.1)$$

Another way to set $\rho_s = 1.0$ g ml^{-1} is to cross-link the crystals with glutaraldehyde (Quiocho & Richards, 1964; Cornick *et al.*, 1973; Matthews, 1985), making the crystals insoluble even in the absence of stabilizing solutes. Once cross-linked, crystals can be transferred to a water solution prior to the density measurement, thereby substituting water for its free solvent. Care must be taken with cross-linking, however. Overnight soaking in 2% glutaraldehyde solutions can substantially increase the crystal density, while destroying its crystalline order (Matthews, 1985). Even 0.5% glutaraldehyde concentrations may change the observed density of some crystals if the exposure is for many hours – which may be necessary to render the crystal completely insoluble. Therefore, the densities observed from cross-linked crystals should be regarded with caution.

If it is necessary to carry out density measurements in an organic solvent gradient, then it is necessary in general to measure the crystal density at more than one free solvent density, since the relative volume fractions of the crystal's components are not known *a priori*. However, if this is a well behaved protein crystal, by setting $\overline{v}_m = 0.74$ ml g^{-1}, $V_M = 2.4$ Å^3 Da^{-1} and $w = 0.25$ g bound water per g protein, one can guess the crystal's volume compartments to be:

$$\varphi_m = 0.51, \qquad \varphi_{fs} = 0.32, \qquad \varphi_{bs} = 0.17,$$

and the mean solvent density to use in (5.2.4.9) would be

$$\rho_s \simeq 0.35 + 0.65\rho_{fs}. \qquad (5.2.7.2)$$

This may give reasonably reliable derivations for n in (5.2.4.9), with just one crystal-density measurement. Over-reliance on parameter estimates, however, can lead to bogus results, and (5.2.7.2) should be used with caution.

This work was supported by the US Department of Energy, Office of Biological and Environmental Research, under contract W31-109-ENG-38.

References

Adair, G. S. & Adair, M. E. (1936). *The densities of protein crystals and the hydration of proteins. Proc. R. Soc. London Ser. B*, **120**, 422–446.

Berman, H. (1939). *A torsion microbalance for the determination of specific gravities of minerals. Am. Mineral.* **24**, 434–440.

Bernal, J. D. & Crowfoot, D. (1934). *Use of the centrifuge in determining the density of small crystals. Nature (London)*, **134**, 809–810.

Bode, W. & Schirmer, T. (1985). *Determination of the protein content of crystals formed by Mastigocladus laminosus C-phycocyanin, Chroomonas spec. phycocyanin-645, and modified human fibrinogen using an* *improved Ficoll density gradient method. Biol. Chem. Hoppe–Seyler*, **366**, 287–295.

Cantor, C. R. & Schimmel, P. R. (1980). *Biophysical Chemistry*. San Francisco: W. H. Freeman & Co.

Coleman, P. M. & Matthews, B. W. (1971). *Symmetry, molecular weight, and crystallographic data for sweet potato β-amylase. J. Mol. Biol.* **60**, 163–168.

Cornick, G., Sigler, P. B. & Ginsberg, H. S. (1973). *Characterization of crystals of type 5 adenovirus hexon. J. Mol. Biol.* **73**, 533–538.

Crick, F. (1957). *X-ray diffraction of protein crystals. Methods Enzymol.* **4**, 127–146.

Edelstein, S. J. & Schachman, H. (1973). *Measurement of partial specific volumes by sedimentation equilibrium in $H_2O - D_2O$ solutions. Methods Enzymol.* **27**, 83–98.

Edsall, J. T. (1953). *Solvation of proteins.* In *The Proteins*, edited by H. Neurath & K. Bailey, Vol. 1, part B, pp. 549–726. New York: Academic Press.

Graubner, H. (1986). *Densitometer for absolute measurements of the temperature dependence of density, partial volumes, and thermal expansivity of solids and liquids. Rev. Sci. Instrum.* **57**, 2817–2826.

Hegerl, R. & Altbauer, A. (1982). *The EM program system. Ultramicroscopy*, **9**, 109–116.

Kiefersauer, R., Stetefeld, J., Gomis-Rüth, F. X., Romão, M. J., Lottspeich, F. & Huber, R. (1996). *Protein-crystal density by volume measurement and amino-acid analysis. J. Appl. Cryst.* **29**, 311–317.

Kratky, O., Leopold, H. & Stabinger, H. (1973). *The determination of the partial specific volume of proteins by the mechanical oscillator technique. Methods Enzymol.* **27**, 98–110.

Kuntz, I. D. & Kaufmann, W. (1974). *Hydration of proteins and polypeptides. Adv. Protein Chem.* **28**, 239–345.

Low, B. W. & Richards, F. M. (1952*a*). *The use of the gradient tube for the determination of crystal densities. J. Am. Chem. Soc.* **74**, 1660–1666.

Low, B. W. & Richards, F. M. (1952*b*). *Determination of protein crystal densities. Nature (London)*, **170**, 412–415.

Matthews, B. W. (1968). *Solvent content of protein crystals. J. Mol. Biol.* **33**, 491–497.

Matthews, B. W. (1974). *Determination of molecular weight from protein crystals. J. Mol. Biol.* **82**, 513–526.

Matthews, B. W. (1977). *Protein crystallography.* In *The Proteins*, edited by H. Neurath & R. L. Hill, pp. 403–590. New York: Academic Press.

Matthews, B. W. (1985). *Determination of protein molecular weight, hydration, and packing from crystal density. Methods Enzymol.* **114**, 176–187.

Midgley, H. G. (1951). *A quick method of determining the density of liquid mixtures. Acta Cryst.* **4**, 565.

Perutz, M. F. (1946). *The solvent content of protein crystals. Trans. Faraday Soc.* **42B**, 187–195.

Phillips, G. N., Lattman, E. E., Cummins, P., Lee, K. Y. & Cohen, C. (1979). *Crystal structure and molecular interactions of tropomyosin. Nature (London)*, **278**, 413–417.

Quiocho, F. A. & Richards, F. M. (1964). *Intermolecular cross linking of a protein in the crystalline state: carboxypeptidase-A. Proc. Natl Acad. Sci. USA*, **52**, 833–839.

Reilly, J. & Rae, W. N. (1954). *Determination of the densities of solids by voluminometry.* In *Physico-Chemical Methods*, Vol. 1, pp. 577–608. New York: van Nostrand.

Richards, F. M. (1954). *A microbalance for the determination of protein crystal densities. Rev. Sci. Instrum.* **24**, 1029–1034.

Richards, F. M. & Lindley, P. F. (2004). *Determination of the density of solids.* In *International Tables for Crystallography*, Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince, ch. 3.2. Dordrecht: Kluwer Academic Publishers.

Russ, J. C. (1990). *Computer-Assisted Microscopy: the Measurement and Analysis of Images.* New York: Plenum Press.

Scanlon, W. J. & Eisenberg, D. (1975). *Solvation of crystalline proteins: theory and its application to available data. J. Mol. Biol.* **98**, 485–502.

Syromyatnikov, F. V. (1935). *The micropycnometric method for the determination of specific gravities of minerals. Am. Mineral.* **20**, 364–370.

Tanford, C. (1961). *Physical Chemistry of Macromolecules.* New York: John Wiley & Sons.

Westbrook, E. M. (1976). *Characterization of a hexagonal crystal form of an enzyme of steroid metabolism, D5–3-ketosteroid isomerase: a new method of crystal density measurement. J. Mol. Biol.* **103**, 659–664.

Westbrook, E. M. (1985). *Crystal density measurements using aqueous Ficoll solutions. Methods Enzymol.* **114**, 187–196.

International Tables for Crystallography (2012). Vol. F, Chapter 6.1, pp. 159–167.

PART 6. RADIATION SOURCES AND OPTICS

Chapter 6.1. X-ray sources

U. W. Arndt

6.1.1. Overview

In this chapter we shall discuss the production of the most suitable X-ray beams for data collection from single crystals of macromolecules. This subject covers the generation of X-rays and the conditioning or selection of the X-ray beam that falls on the sample with regard to intensity, cross section, degree of parallelism and spectral composition. The conclusions drawn do not necessarily apply to smaller-unit-cell crystals or to noncrystalline samples.

6.1.2. Generation of X-rays

X-rays are generated by the interaction of charged particles with an electromagnetic field. There are four sources of interest to the crystallographer.

(1) The bombardment of a target by electrons in a vacuum tube produces a continuous ('white') X-ray spectrum, called *Bremsstrahlung*, which is accompanied by a number of discrete spectral lines characteristic of the target material. The most common target material is copper, and the most frequently employed X-ray line is the copper $K\alpha$ doublet with a mean wavelength of 1.542 Å. X-ray tubes are described in some detail in Chapter 4.2 of *IT* C (2004). We shall consider only the most important points in X-ray tube design here.

(2) Synchrotron radiation is produced by relativistic electrons in orbital motion. This is the subject of Part 8.

(3) The decay of natural or artificial radioisotopes is often accompanied by the emission of X-rays. Radioactive sources are often used for the testing and calibration of X-ray detectors. For our purposes, the most important source is made from ^{55}Fe, which has a half-life of 2.6 years and produces Mn $K\alpha$ X-rays with an energy of 5.90 keV.

(4) Ultra-short pulses of X-rays are generated in plasmas produced by the bombardment of targets by high-intensity sub-picosecond laser pulses (*e.g.* Forsyth & Frankel, 1984). In earlier work, the maximum pulse repetition frequency was much less than 1 Hz, but picosecond pulses at more than 1 Hz are now being achieved with µm-size sources. The time-averaged X-ray intensities from these sources are very low, so their application will probably remain limited to time-resolved studies (Kleffer *et al.*, 1993).

X-rays also arise in the form of channelling radiation resulting from the bombardment of crystals, such as diamonds, by electrons with energies of several MeV from a linear accelerator (Genz *et al.*, 1990) and in the form of transition radiation when multiple-foil targets are bombarded by electrons in the range 100–500 MeV (*e.g.* Piestrup *et al.*, 1991). It will be some time before these new sources can compete with the older methods for routine data collection.

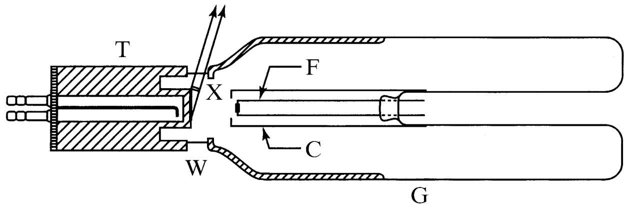


Figure 6.1.2.1
Section through a sealed X-ray tube. G, glass envelope; F, filament leads (at negative high voltage); C, focusing cup; T, target (at ground potential); W, one of four beryllium windows. The electron beam forms a line on the target, which is viewed at a small take-off angle to form a foreshortened effective source X.

6.1.2.1. Stationary-target X-ray tubes

A section through a permanently evacuated, sealed X-ray tube is shown in Fig. 6.1.2.1 The tube has a spirally wound tungsten filament, F, placed immediately behind a slot in the focusing cup, C, and a water-cooled target or anode, T, approximately 10 mm from the surface of C. The filament–focusing-cup assembly is at a negative voltage of between 30 and 50 kV, and the target is at ground potential. The electron beam strikes the target in a focal line, which acts as a line source of X-rays. There are usually two pairs of X-ray windows, W, through which the source is viewed at a small angle to the target surface, thus producing a foreshortened effective source, X, which is approximately square in one plane and a narrow line in the other. Focus dimensions on the target and maximum recommended power loading are shown for a number of standard inserts in Table 6.1.2.1. None of these are ideal for macromolecular crystallography. The assembly of a cathode, anode and windows – the tube insert – is inserted in a shock- and radiation-proof shield which is fixed to the table. Attached to the shield are X-ray shutters and filters, and sometimes brackets for bolting on X-ray cameras. A high-voltage connection is made to the tube by means of a flexible, shielded, shock-proof cable; nowadays, this high voltage is almost invariably full-wave rectified and smoothed DC.

6.1.2.2. Rotating-anode X-ray tubes

The sealed tubes described above are convenient and require little maintenance, but their power dissipation, and thus their X-ray output, is limited. For macromolecular crystallography, the most commonly used tubes are continuously pumped, demountable tubes with water-cooled rotating targets [see the

Table 6.1.2.1
Standard X-ray tube inserts

Focus size on target (mm × mm)	Recommended power loading (kW)
8 × 0.15	0.8
8 × 0.4	1.5
10 × 1.0	2.0
12 × 2.0	2.7

Copyright © 2012 International Union of Crystallography

reviews by Yoshimatsu & Kozaki (1977) and Phillips (1985)]. At present, these tubes mostly employ ferro-fluidic vacuum shaft seals (Bailey, 1978), which have an operational life of several thousand hours before they need replacement. The need for a beam with a small cross fire calls for a focal spot preferably not larger than 0.15 × 0.15 mm. This is usually achieved by focusing the electrons on the target to a line 0.15 mm wide and 1.5 mm long (in the direction parallel to the rotation axis of the target). The line is then viewed at an angle of 5.7° to give a 10:1 foreshortening. Foci down to this size can be produced on a target mounted close to an electron gun. For smaller focal spots, such as those of the microfocus tube described below, it is necessary to employ an electron lens (which may be magnetic or electrostatic) to produce on the target a demagnified image of the electron cross-over, which is close to the grid of the tube cathode.

The maximum power that can be dissipated in the target without damaging the surface has been discussed by Müller (1929, 1931), Oosterkamp (1948*a*,*b*,*c*), and Ishimura *et al.* (1957). The later calculations are in adequate agreement with Müller's results, from which the power W for a copper target is given by

$$W = 26.4\, f_1 (f_2 v)^{1/2}.$$

Here, W is in watts, f_1 and f_2 are the length and the width of the focal line in mm, and v is the linear speed of the target surface in mm s^{-1}; it is assumed that the surface temperature of the target reaches 600 °C, well below the melting point of copper (1083 °C). Thus, for a focal spot 1.5 × 0.15 mm and for an 89 mm-diameter target rotating at 6000 revolutions min^{-1} (v = 28 000 mm s^{-1}), Müller's formula gives a maximum permissible power loading of 2.5 kW or 57 mA at 45 kV. This agrees well with the experimentally determined loading limit.

Green & Cosslett (1968) have made extensive measurements of the efficiency of the production of characteristic radiation for a number of targets and for a range of electron accelerating voltages. Their results have been verified by many subsequent investigators. For a copper target, they found that the number of $K\alpha$ photons emitted per unit solid angle per incident electron is given by

$$N/4\pi = 6.4 \times 10^{-5} [E/E_k - 1]^{1.63},$$

where E is the tube voltage in kV and E_k = 8.9 keV is the K excitation voltage.

Accordingly, the number of $K\alpha$ photons generated per second per steradian per mA of tube current is 1.05 × 10^{12} at 25 kV and 4.84 × 10^{12} at 50 kV.

Of the generated photons, only a fraction, usually denoted by $f(\chi)$ (Green, 1963), emerges from the target as a result of X-ray absorption in the target. $f(\chi)$ decreases with increasing tube voltage and with decreasing take-off angle. It has a value of about 0.5 for E = 50 kV and for a take-off angle of 5°.

The X-ray beam is further attenuated by absorption in the tube window (~80% transmission), by the air path between the tube and the sample, and by any β-filters which may be used.

In a typical diffractometer or image-plate arrangement where no beam conditioning other than a β-filter is employed, the sample may be 300 mm from the tube focus and the limiting aperture at that point might have a diameter of 0.3 mm, so that the full-angle cross fire at the sample is 1.0 × 10^{-3} rad. The solid angle subtended by the limiting aperture at the source is 7.9 × 10^{-7} steradians. At 50 kV and 60 mA, the X-ray flux through the sample will be approximately 4.5 × 10^{7} photons s^{-1}. These figures are approximately confirmed by unpublished experimental measurements by Arndt & Mancia and by Faruqi & Leslie. It is interesting to note that the power in this photon flux is 5.8 × 10^{-8} W, which is a fraction of 2 × 10^{-11} of the power loading of the X-ray tube target.

Instead of simple aperture collimation, one of the types of focusing collimators described in Section 6.1.4.1 below may be used. They collect a somewhat larger solid angle of radiation from the target of a conventional X-ray source than does a simple collimator and some produce a higher intensity at the sample.

6.1.2.3. *Microfocus X-ray tubes*

Standard sealed X-ray tubes with a stationary target deliver a collimated intensity to the sample which is insufficient for most applications in macromolecular crystallography. These tubes have foreshortened foci between 0.4 and 2 mm^2 which do not lend themselves to efficient collimation by means of focusing mirrors or monochromators without introducing a cross fire in the beam that is too large for our purposes.

The situation is different with microfocus tubes, which are discussed in Section 6.1.4.2. Here, a relatively large solid angle of collection can make up for the lower power dissipation which results from the small electron focus.

6.1.2.4. *Synchrotron-radiation sources*

Charged particles with energy E and mass m moving in a circular orbit of radius R at a constant speed v radiate a power, P, into a solid angle of 4π, where

$$P = 88.47 E^4 I/R,$$

where E is in GeV, I is the circulating electron or positron current in ampères and R is in metres. Thus, for example, in a bending-magnet beam line at the ESRF, Grenoble, France, R = 20 m, and at 5 GeV and 200 mA, P = 554 kW.

For relativistic electrons, the electromagnetic radiation is compressed into a fan-shaped beam tangential to the orbit, with a

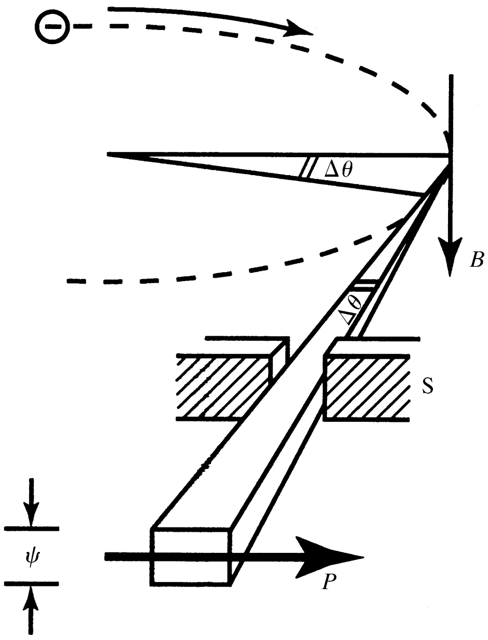


Figure 6.1.2.2
Synchrotron radiation emitted by a relativistic electron travelling in a curved trajectory. B is the magnetic field perpendicular to the plane of the electron orbit; ψ is the natural opening angle in the vertical plane; P is the direction of polarization. The slit, S, defines the length of the arc of angle, $\Delta\theta$, from which the radiation is taken. From Buras & Tazzari (1984).

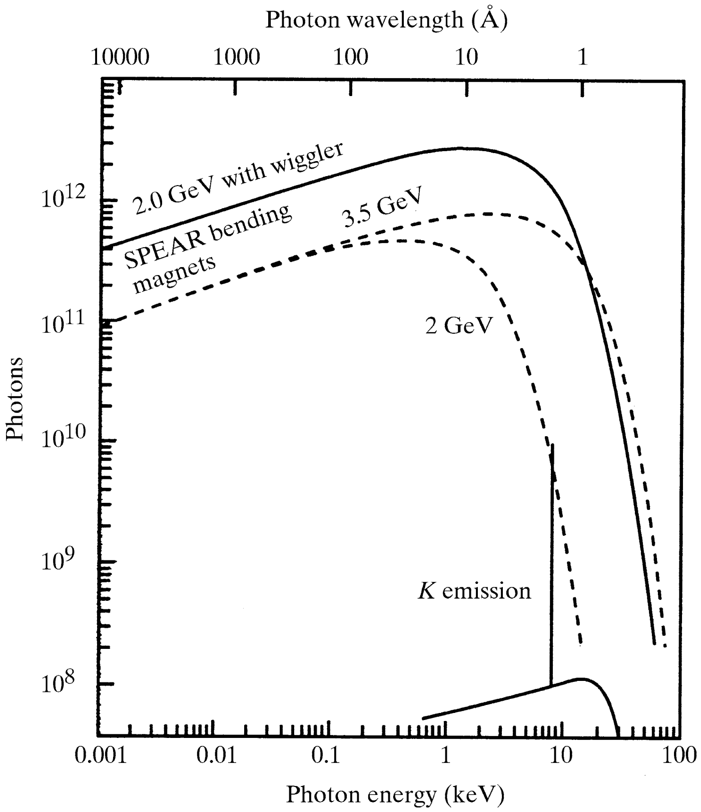


Figure 6.1.2.3
Comparison of the spectra from the storage ring SPEAR in photons s^{-1} mA^{-1} $mrad^{-1}$ per 1% band pass (1978 performance) and a rotating-anode X-ray generator, showing the Cu *K* emission line and the *Bremsstrahlung*. Reproduced with permission from Nagel (1980). Copyright (1980) New York Academy of Sciences.

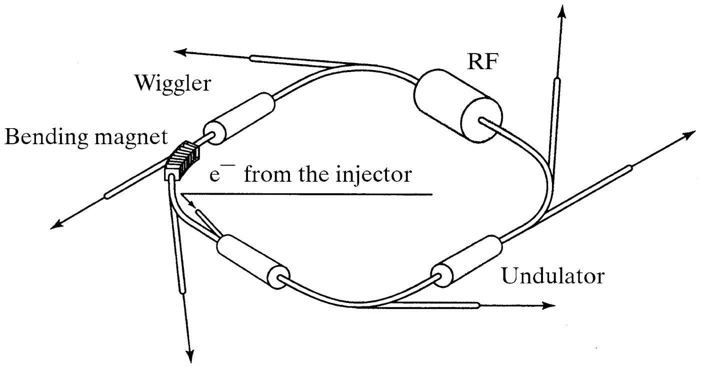


Figure 6.1.2.4
Main components of a dedicated electron storage-ring synchrotron-radiation source. For clarity, only one bending magnet is shown. From Buras & Tazzari (1984).

vertical opening angle $\Psi \simeq mc^2/E$, *i.e.* 0.1 mrad for $E = 5$ GeV (Fig. 6.1.2.2). This fan rotates with the circulating electrons; if the ring is filled with n bunches of electrons, a stationary observer will see n flashes of radiation every $2\pi R/c$ s, the duration of each flash being less than 1 ns.

The spectral distribution of synchrotron radiation extends from the infrared to the X-ray region; Schwinger (1949) gives the instantaneous power radiated by a monoenergetic electron in a circular motion per unit wavelength interval as a function of wavelength (Winick, 1980). An important parameter specifying the distribution is the critical wavelength, λ_c: half the total power radiated, but only ~9% of the total number of photons, is at $\lambda < \lambda_c$ (Fig. 6.1.2.6). λ_c (in Å) is given by

$$\lambda_c = 18.64/(BE^2),$$

where B (= $3.34E/R$) is the magnetic bending field in T, E is in GeV and R is in metres.

The orbit of the particle can be maintained only if the energy lost, in the form of electromagnetic radiation, is constantly replenished. In an electron synchrotron or in a storage ring, the circulating particles are electrons or positrons maintained in a closed orbit by a magnetic field; their energy is supplied or restored by means of an oscillating radiofrequency (RF) electric field at one or more places in the orbit. In a synchrotron designed for nuclear-physics experiments, the circulating particles are injected from a linear accelerator, accelerated up to full energy by the RF field, and then deflected onto a target with a cycle frequency of about 50 Hz. The synchrotron radiation is thus produced in the form of pulses of this frequency. A storage ring, on the other hand, is filled with electrons or positrons, and after acceleration the particle energy is maintained by the RF field; ideally, the current circulates for many hours and decays only as a result of collisions with remaining gas molecules. At present, only storage rings are used as sources of synchrotron radiation, and many of these are dedicated entirely to the production of radiation: they are not used at all, or are used only for limited periods, for nuclear-physics collision experiments.

Synchrotron radiation is highly polarized. In an ideal ring, where all electrons are parallel to one another in a central orbit, the radiation in the orbital plane is linearly polarized with the electric vector lying in this plane. Outside the plane, the radiation is elliptically polarized.

In practice, the electron path in a storage ring is not a circle. The 'ring' consists of an alternation of straight sections and bending magnets. Beam lines are installed at the bending magnets and at the insertion devices.

These insertion devices with a zero magnetic field integral, *i.e.* wigglers and undulators, may be inserted in the straight sections (Fig. 6.1.2.4). A wiggler consists of one or more dipole magnets with alternating magnetic field directions aligned transverse to the orbit. The critical wavelength can thus be shifted towards shorter values because the bending radius can be decreased over a short section, especially when superconducting magnets are used. Such a device is called a wavelength shifter. If it has N dipoles, the radiation from the different poles is added to give an N-fold increase in intensity. Wigglers can be horizontal or vertical. In a wiggler, the maximum divergence, 2α, of the electron beam is much larger than ψ, the vertical aperture of the radiation cone in the spectral region of interest (Fig. 6.1.2.2). If $2\alpha \ll \psi$, and if, in addition, the magnet poles of a multipole device have a short period, then the device becomes an undulator; interference takes place between the radiation of wavelength λ_0 emitted at two points λ_0 apart on the electron trajectory (Fig. 6.1.2.5). The spectrum at an angle θ to the axis observed through a pinhole consists of a single spectral line and its harmonics with wavelengths

$$\lambda_i = i^{-1}\lambda_0[(E/mc^2)^{-2} + \alpha^2/2 + \theta^2]/2$$

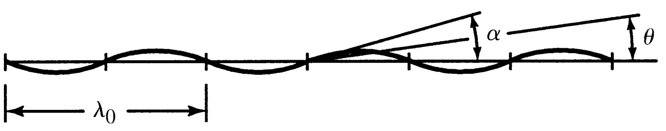


Figure 6.1.2.5
Electron trajectory within a multipole wiggler or undulator. λ_0 is the spatial period, α is the maximum deflection angle and θ is the observation angle. From Buras & Tazzari (1984).

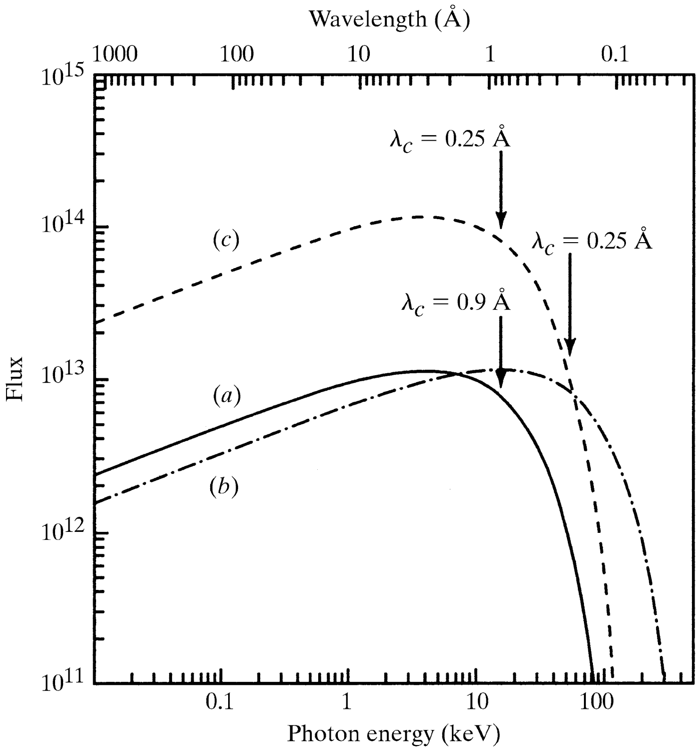


Figure 6.1.2.6

Spectral distribution and critical wavelengths for (*a*) a dipole magnet, (*b*) a wavelength shifter and (*c*) a multipole wiggler at the ESRF. From Buras & Tazzari (1984).

(Hofmann, 1978). Typically, the bandwidth of the lines, $\delta\lambda/\lambda$, is ~0.01 to 0.1, and the photon flux per unit bandwidth from the undulator is many orders of magnitude greater than that from a bending magnet. Undulators at the ESRF have a fundamental wavelength of less than 1 Å.

The spectra for a bending magnet and a wiggler are compared with that from a copper-target rotating-anode tube in Fig. 6.1.2.3.

6.1.3. Properties of the X-ray beam

We must now consider the properties of the X-ray beam necessary for the gathering of intensity data from single crystals of biological macromolecules. The properties of the beam with which we are concerned are:

(*a*) the size of the beam appropriate for the sample dimensions;
(*b*) the X-ray wavelength and its spectral purity;
(*c*) the intensity in photons s^{-1};
(*d*) the cross fire, that is, the maximum angle between rays in the beam;
(*e*) the temporal structure of the beam, that is, its stability or constancy, and for generators other than X-ray tubes, the duration and frequency of intensity pulses.

These properties cannot be considered in isolation since the requirements depend on the particular crystal under investigation (size, unit-cell dimensions, mosaic spread and resistance to radiation damage), on the geometry of the X-ray camera or diffractometer and on the detector used.

6.1.3.1. Beam size

The best signal-to-noise ratio in the diffraction pattern is secured when the sample crystal is just bathed in the X-ray beam, which is often taken to be about 0.2 to 0.3 mm in diameter. Unfortunately, many crystals are plate- or needle-shaped and present a greatly varying aspect to the beam. To date, no-one has described data-collection instruments in which the incident-beam dimension is changed automatically as the crystal is rotated; the next best thing is a versatile collimation system that makes use of interchangeable beam-limiting apertures.

6.1.3.2. X-ray wavelength

For X-ray tube sources, the main component of the beam is the characteristic radiation of the tube target. The vast majority of macromolecular structure determinations have been carried out with copper $K\alpha$ X-rays of wavelength 1.54 Å. These are reasonably well matched to the linear absorption coefficients of biological materials. Diffractometers and cameras are usually designed to permit data collection out to Bragg angles of about 30°, that is, to a minimum spacing of 1.54 Å, which is a convenient limit.

The next shortest, useful characteristic X-rays are, in practice, those from a molybdenum target (0.71 Å), but are rarely used in macromolecular crystallography.

The advantages of shorter wavelengths are a reduced absorption correction, smaller angles of incidence on the film, image plate or area detector, and, probably, a slightly smaller amount of radiation damage for a given intensity of the diffraction pattern. The disadvantage is a lower diffracted intensity, which is approximately proportional to the square of the wavelength. Crystal monochromators and specularly reflecting X-ray mirrors have a lower reflectivity for shorter wavelengths; most X-ray detectors, other than image plates and scintillation counters, are less efficient for harder X-rays (see Part 7).

At synchrotron beam lines where there is no shortage of X-ray intensity, it is now customary to select X-ray wavelengths of about 1 Å for routine data collection. Here, of course, it is possible to choose optimum wavelengths for anomalous-dispersion phasing experiments. This possibility is one of the major advantages of synchrotron radiation. The selection of a narrow wavelength band from the white radiation continuum (*Bremsstrahlung*) of an X-ray tube by means of crystal monochromators is not of practical importance: a tungsten-target X-ray tube operated at 80 kV produces about 1×10^{-5} 8 keV photons per steradian per electron incident on the target within a wavelength band, $\delta\lambda/\lambda$, of 10^{-3}; a copper-target X-ray tube at 40 kV produces about 5×10^{-4} $K\alpha$ photons per steradian per electron, that is, about 50 times more X-rays.

6.1.3.3. Spectral composition

Any X-rays outside the wavelength band used for generating the desired X-ray pattern contribute to the radiation damage of the sample and to the X-ray background. In the interests of resolving neighbouring diffraction spots in the pattern, one would require the wavelength spread, $\delta\lambda/\lambda$, in the incident radiation to be less than 5×10^{-3}. For the Cu $K\alpha$ doublet $(\lambda_{\alpha_2} - \lambda_{\alpha_1})/\lambda \simeq 2.5 \times 10^{-3}$ and the doublet nature of the line usually does not matter. On the other hand, the value of $(\lambda_\alpha - \lambda_\beta)/\lambda$ is 0.1, so the $K\beta$ component must be eliminated by means of a β-filter (a 0.15 mm-thick nickel foil for copper radiation) or by reflection from a crystal monochromator to avoid the appearance of separate $K\beta$ diffraction spots. The dispersion produced by a crystal monochromator is small enough to be ignored in most applications.

In synchrotron beam lines, the bandpass is usually determined by the divergence of the beam and is of the order of 10^{-4}. This is a smaller bandpass than is required for most purposes, and intensity can be gained by widening the bandpass by the use of an asymmetric-cut monochromator in spatial expansion geometry (Nave *et al.*, 1995; Kohra *et al.*, 1978). The intensity outside the monochromator bandpass is usually totally negligible.

6.1.3.4. *Intensity*

The intensity of the primary X-ray beam should be such as to allow data collection in a reasonably short time; increased speed is one of the main factors which has led to the popularity of synchrotron-radiation data collection as compared to data collection using conventional sources. Moreover, the radiation damage to the sample per unit incident dose is smaller at high intensities. This does not mean that ever more intense beams are necessary for today's protein-crystallography problems; very often, the speed of data collection is limited by the read-out time of the detector; the counting-rate capabilities of present-day X-ray detectors make it impossible to use in full the intensities available at some beam lines. With the widespread use of cryocrystallographic methods (Part 10), radiation damage is no longer as severe a problem as it once was.

No doubt, the day will come when available intensities will be so high that instantaneous structure determination will become a possibility, but this will require major advances in X-ray detectors, probably in the form of the development of large pixel detectors (*e.g.* Beuville *et al.*, 1997).

There is still some scope for increasing the intensity of X-ray beams from conventional sources, which offer the advantage of making measurements in the local laboratory instead of at some remote central facility.

6.1.3.5. *Cross fire*

The cross fire is defined as the angle or the half-angle between extreme rays in the beam incident on a given point of the sample. In the absence of focusing elements, such as specularly reflecting mirrors, the X-ray beam diverges from the source. A diverging beam can be turned into a converging one by reflection from a curved mirror or crystal. A crystal with constant lattice spacing can change the sign, but not the magnitude, of the angle between rays, whatever the curvature of the crystal; the deviation produced by a reflection anywhere on the surface of the crystal must always be twice the Bragg angle. It is possible to change a divergent beam into a convergent one with a different cross fire by specular reflection at a mirror or by means of a crystal in which the lattice spacing varies from point to point along the length of the plate. Such variable-spacing reflectors may be either artificial-crystal multilayers (Schuster & Göbel, 1997) or, less commonly, natural crystals whose spacing is modified by variable doping or by a temperature gradient along the crystal plate (Smither, 1982). In the neutron-scattering community, graded-spacing multilayer monochromators are usually referred to as 'supermirrors' (see Section 6.2.1.3.3 in Chapter 6.2).

The construction of curved mirrors and curved-crystal monochromators is discussed below.

As a general rule, the better the collimation of the incident X-ray beam, that is, the smaller the cross fire and the more nearly parallel the beam, the cleaner the diffraction pattern and the lower the background. Synchrotron-radiation sources permit a certain prodigality in X-ray intensity and the beams from them are thus usually better collimated than beams from conventional sources.

The useful degree of collimation depends on the crystal under investigation. The aim in the design of an instrument for data collection from single crystals must be to make the widths of the angular profiles of the reflection small, but these widths cannot be reduced beyond the rocking-curve width determined by the mosaic spread of the sample, however small the cross fire. The mosaic spread of a typical protein crystal is often quoted as being about 10^{-3} rad or about 3.4 minutes of arc. However, there are many crystals with much larger mosaicities; for such samples, the intensity of the X-ray beam, expressed as the total number of photons which strike the crystal, can be increased by permitting a larger cross fire.

The mosaic spread must be understood as the angle between individual domains of the mosaic crystal. These domains may be as large as 100 µm, that is, they may have dimensions not so very much smaller than those of the macroscopic crystal. The individual rocking-curve widths may be as small as 10 seconds of arc (50 µrad). Fourme *et al.* (1995) have discussed the implications of this degree of perfection if the collimation is improved to a stage where it can be exploited.

The way in which the cross fire influences the angular widths of the reflections depends on the instrument geometry. In a single-counter four-circle diffractometer, all reflections are brought onto the equator and the crystal is rotated about an axis perpendicular to the equatorial plane. The cross fire should, therefore, be small parallel to the equatorial plane, *i.e.* usually in the horizontal plane. The cross fire in the plane containing the rotation axis affects the angular width of the reflections much less, and it could thus be made larger in the interest of a high intensity.

The situation is different if the diffractometer is fitted with an electronic area detector, such as a CCD or other TV detector or a multi-wire proportional chamber (see Part 7). Here, the widths of reflections in upper levels are affected by the cross fire in the plane containing the crystal rotation axis, and the divergence or convergence of the beam in this plane should also be kept small.

With recording on photographic film or image plates, each exposure, or 'shot' or 'frame', corresponds to a crystal rotation that is usually many times larger than the angular width of a reflection. It is then less important to keep the cross fire small in the plane perpendicular to the rotation axis of the crystal.

In many collimation arrangements, the cross fire can be chosen independently in the two planes. In the absence of monochromators or mirrors, the cross fire is determined by beam apertures, which can be rectangular slits; it is, of course, simpler to employ circular holes, which give the same cross fire in both planes.

6.1.3.6. *Beam stability*

The synchrotron beam decays steadily after each filling of the ring as the number of stored positrons or electrons decays. Even with an X-ray tube operated from voltage- and current-stabilized supplies, the X-ray intensity changes with time as a result of contamination and roughening of the target surface. It is, thus, highly desirable to have a method of monitoring the beam incident on the sample, for example, by means of an ionization chamber built into the collimator (Arndt & Stubbings, 1988).

It should be noted that when the collimator contains focusing elements, the intensity at the sample can vary by several hundred per cent, depending on the exact alignment of the focusing mirrors or crystals and on the exact dimensions of the electron focus on the tube target.

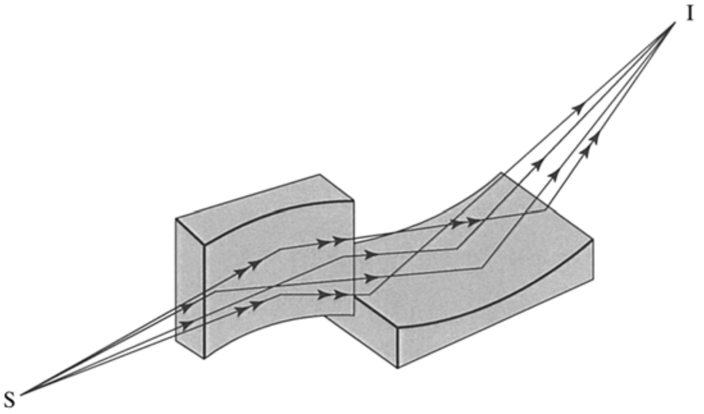


Figure 6.1.4.1
Production of a point focus by successive reflections at two orthogonal curved mirrors. Arrangement due to Kirkpatrick & Baez (1948) and to Franks (1955).

Intensity changes can be caused by mechanical movement of collimating components. Among these may be such unsuspected effects as flexing of the target surface with changes in cooling-water pressure.

The response of an incident-beam monitor may itself vary as a result of changes in temperature, barometric pressure, or humidity.

Synchrotron radiation from storage rings has a regular time-dependent modulation brought about by the rate of passage of bunches of electrons or positrons in the ring. For the great majority of measurements, this time structure has no effect, but at very high intensities, the counting losses are greater than they would be from a steady source.

6.1.4. Beam conditioning

The primary X-ray beam from the source is conditioned by a variety of devices, such as filters, mirrors and monochromators, to produce the appropriate properties for the beam incident on the sample.

6.1.4.1. X-ray mirrors

It is usually necessary to focus the X-ray beam in two orthogonal directions. This can be achieved either by means of one mirror with curvatures in two orthogonal planes or by two successive reflections from two mirrors which are curved in one plane and planar in the other; the two planes of curvature must be at right angles to one another. In the arrangement adopted by Kirkpatrick & Baez (1948) and by Franks (1955), the two mirrors lie one behind the other (Fig. 6.1.4.1) and thus produce a different degree of collimation in the two planes. Instead of this tandem arrangement, the mirrors can lie side-by-side, as proposed by Montel (1957), to form what the author calls a ‘catamegonic roof’ (Fig. 6.1.4.2). The mirrors are then best made from thicker material, and the reflecting surfaces are ground to the appropriate curvature. The same arrangement has been used by Osmic Inc. (1998) for their Confocal Max-Flux Optics, in which the curved surfaces are coated with graded-spacing multilayers.

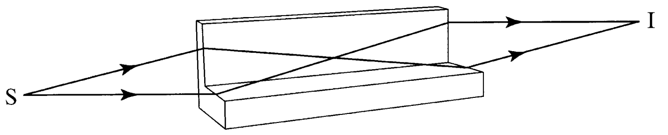


Figure 6.1.4.2
The ‘catamegonic’ arrangement of Montel (1957), in which two confocal mirrors with orthogonal curvatures lie side-by-side.

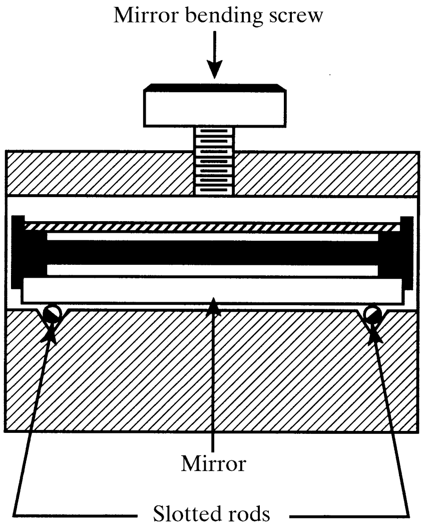


Figure 6.1.4.3
Mirror bender (after Franks, 1955). The force exerted by the screw produces two equal couples which bend the mirror into a circular arc. The slotted rods act as pivots and also as beam-defining slits.

Flat mirror plates can be bent elastically to a desired curvature by applying appropriate couples. Fig. 6.1.4.3 shows the bending method adopted by Franks (1955). A cylindrical curvature results from a symmetrical arrangement that produces equal couples at both ends. With appropriate unequal couples applied at the two ends of the plate, the curvature can be made parabolic or elliptical. Precision elliptical mirrors have been produced by Padmore *et al.* (1997); unequal couples are applied in this way. Cylindrically curved mirrors can be produced by applying a force at the tip of a triangular plate whose base is firmly anchored (Fig. 6.1.4.4). Lemonnier *et al.* (1978) first used this method for making curved-crystal monochromators. Milch (1983) described X-ray mirrors made in this way; the effect of the linear increase of the bending moment along the plate is compensated by the linear increase of the plate section so that the curvature is constant. An elliptical or a parabolic curvature results if either the width or the thickness of the plate is made to vary in an appropriate way along the length of the plate. Arndt, Long & Duncumb (1998) described a monolithic mirror-bending block in which the mirror plates are

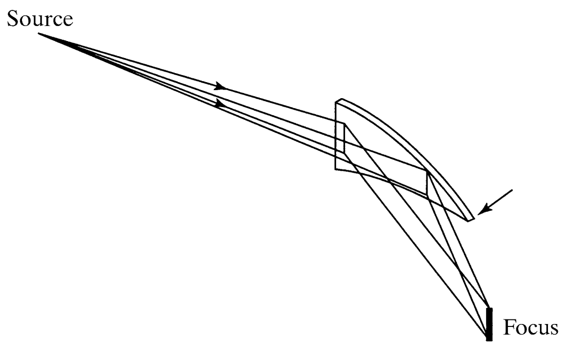


Figure 6.1.4.4
Triangular mirror bender as described by Lemonnier *et al.* (1978) for crystal plates and by Milch (1983) for glass mirrors. The base of the triangular plate is clamped and the bending force is applied at the apex along the arrow.

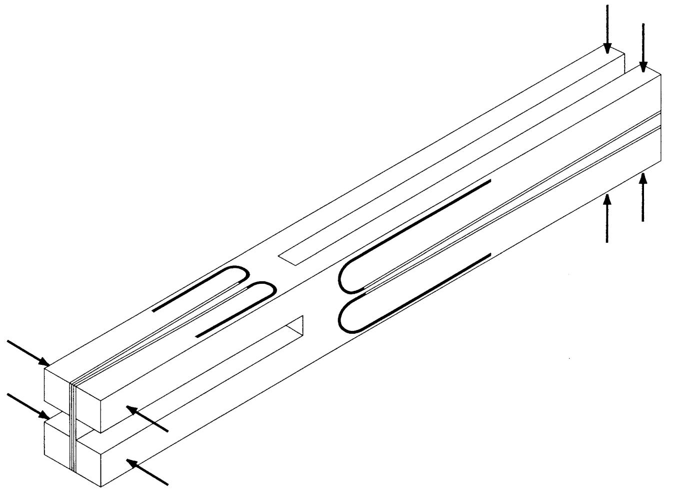


Figure 6.1.4.5
Mirror holder with machined slots for two orthogonal pairs of curved mirrors (after Arndt, Duncumb *et al.*, 1998).

inserted into slots cut to an elliptical curvature by ion-beam machining. The solid angle of collection is made four times larger than for a two-mirror arrangement by providing a pair of horizontal mirrors and a pair of vertical mirrors in tandem in one block (Fig. 6.1.4.5).

Mirror plates for these benders are usually made from highly polished glass, quartz, or silicon plates which are coated with nickel, gold, or iridium.

Mirrors for synchrotron beam lines that focus the radiation in the vertical plane are most often ground and polished to the correct shape, rather than bent elastically. Much longer mirrors can be made in this way.

The collecting efficiency of specularly reflecting mirrors depends on the reflectivity of the surface and on the solid angle of collection; this, in turn, is a function of the maximum glancing angle of incidence, which is the critical angle for total external reflection, θ_c. For X-rays of wavelength λ, measured in Å,

$$\theta_c \simeq 2.32\times10^{-3}(Z\rho/A)^{1/2}\lambda,$$

where Z is the atomic number, A is the atomic mass and ρ is the specific gravity of the reflecting surface.

Thus, for Cu $K\alpha$ radiation and a gold surface, $\theta_c \simeq 10$ mrad. The reflectivity of the mirror surface is strongly dependent on the surface roughness; for the reflectivity to be more than 50%, the r.m.s. roughness must not exceed 10 Å.

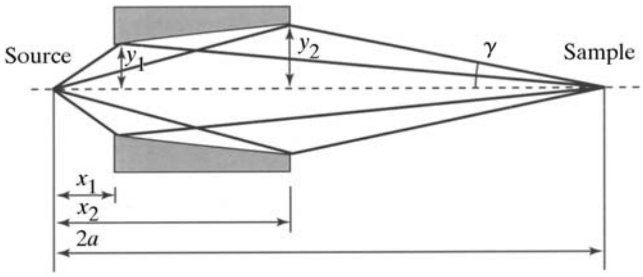


Figure 6.1.4.6
Ellipsoidal mirror for use with a microfocus X-ray tube, where x_1 is ~15 mm. The major axis, $2a$, may be up to 600 mm, whereas the exit aperture, $2y_2$, lies in the region 0.8–1.4 mm. The angle γ determines the cross fire on the sample and is less than 1 rad.

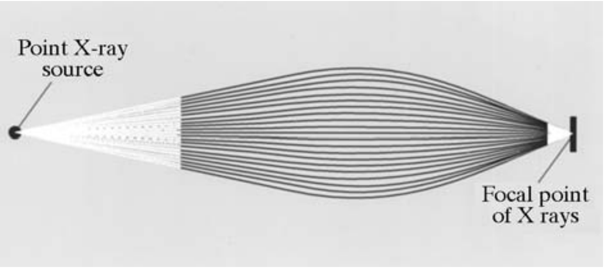


Figure 6.1.4.7
A polycapillary collimator (after Bly & Gibson, 1996).

It is not possible to design a reflecting collimator with a planar angle of collection greater than about $3\theta_c$. For the shorter wavelengths, in particular, variable-spacing multilayer mirrors (Schuster & Göbel, 1997) hold considerable promise. If the spacing at the upstream end of the mirror is 30 Å, the largest angles of incidence will be 26 and 17 mrad for 1.54 and 1.0 Å X-rays, respectively. By comparison, the critical angles at a gold surface for these radiations are 10 and 6.5 mrad, respectively.

6.1.4.2. Focusing collimators for microfocus sources

In most arrangements that include conventional X-ray tubes, the planar angle of collection is very small. A more efficient use is always made of the radiation from the target by a focusing collimator, which forms an image of the source on the sample (Fig. 6.1.4.6). The angle of collection should be as large as possible, while the cross fire, *i.e.* the angle of convergence, is kept small, say, at about 10^{-3} rad. It is possible to design focusing collimators based on gold-surfaced toroids of revolution (Elliott, 1965), which afford a planar angle of collection of about three times the critical angle for total external reflection, that is, about 30×10^{-3} rad.

Consequently, the mirror should *magnify* about 30 times, and if the image diameter, determined by a typical sample size, is to be 300 µm, the size of the focus should be about 10 µm. The solid angle of collection of such an imaging toroid is about 8×10^{-4} steradians, that is, more than 1000 times greater than the solid angle of a simple non-imaging collimator. The averaged mirror reflectivity achieved at present is about 0.3, so the microfocus tube and toroidal mirror combination produces a similar intensity at the sample as the conventional tube with a non-focusing collimator at about 300 times the power. Future increases of the reflectivity are likely as the surface roughness of the mirrors is improved.

A suitable microfocus tube has been described by Arndt, Long & Duncumb (1998); mirrors used with this tube were discussed by Arndt, Duncumb *et al.* (1998). The tube design allows the distance between the source and the mirror to be as little as 10 mm in order to achieve the necessary magnification without making the distance between the tube and the sample inconveniently long.

6.1.4.3. Other focusing collimators

There has been very active development in recent years of tapering capillaries for focusing X-rays, either as individual capillaries (see the review by Bilderback *et al.*, 1994), or in the form of multicapillary bundles. The latter were first described by Kumakhov & Komarov (1990); since then, they have undergone great improvements in the form of fused bundles (Bly & Gibson, 1996) (Fig. 6.1.4.7). Single capillaries have found the greatest use as X-ray concentrators, where a larger-diameter beam of X-rays enters the large end of a tapered capillary and is concentrated to a diameter of a few µm. Fused polycapillary bundles have been employed as focusing collimators for protein crystallography (MacDonald *et al.*, 1999). Both types of capillary optics are usually designed as multi-bounce devices, in which the X-rays undergo several, or many, reflections at the walls of the capillary; consequently the cross-fire half-angle at the output end has a value about equal to the critical angle for reflection at a glass surface or, perhaps, 4 mrad. This is sometimes too great for producing diffraction patterns with an optimum signal-to-background ratio.

Other methods of focusing X-rays, such as zone plates (Kirz, 1974) and refractive optics, are being investigated, but at present none of them can compare with toroidal reflectors for data collection from single crystals of macromolecules.

6.1.4.4. *Crystal monochromators*

When the X-rays from the tube target are specularly reflected by a mirror, the spectrum is cut off for X-rays below the shortest wavelength for which the critical angle is equal to the smallest angle of incidence on the mirror. For a typical mirror designed for Cu $K\alpha$ radiation, this cutoff wavelength might be about 0.75 Å, and the harder X-rays can be further attenuated by a β-filter. Of course, the more nearly monochromatic the radiation falling on the sample, the lower the radiation damage and the higher the spot-to-background ratio in the recorded patterns.

White radiation is almost completely eliminated by reflecting the primary X-ray beam using a natural or artificial (multilayer) crystal. The most commonly used type of plane monochromator for macromolecular crystallography is a single crystal of graphite. This material (HOPG, or highly ordered pyrolytic graphite) has a relatively large mosaic spread, typically about 0.4°, and it cannot separate the $K\alpha$ doublet. This separation is essential in most small-molecule investigations, but is unnecessary for macromolecular crystals, which rarely diffract beyond 1.5 Å, and disadvantageous where a high intensity of the beam reflected by the monochromator is the main consideration.

The intensity of the diffraction pattern obtained with a graphite monochromator is only about two or three times lower than that resulting from a β-filtered pinhole-collimated beam. The situation is different at synchrotron beam lines, which must incorporate a monochromator in order to select the desired X-ray energy band. Curved focusing crystals collect X-rays over a relatively large horizontal angular range and thus produce a beam with a horizontal convergence angle of up to several milliradians. Much more nearly parallel beams are produced by reflection at several crystals in tandem, often in the form of monolithic channel-cut monochromators. In present-day storage rings, the power density at the first optical element is of the order of 10 W mm^{-2} at wiggler and undulator beam lines. This amount of power can be dissipated by careful design of water-cooling channels (Quintana & Hart, 1995; van Silfhout, 1998). In addition, the monochromator crystal, usually of silicon or germanium, may be profiled to minimize distortions as a result of thermal stresses.

The next generation of insertion devices will subject the optical elements to loads of several hundred W mm^{-2}. Possible engineering solutions to the very severe heat-loading problem include the use of diamond crystals as reflecting elements. This material has a very high thermal conductivity, especially at low temperatures.

References

Arndt, U. W., Duncumb, P., Long, J. V. P., Pina, L. & Inneman, A. (1998). *Focusing mirrors for use with microfocus X-ray tubes. J. Appl. Cryst.* **31**, 733.

Arndt, U. W., Long, J. V. P. & Duncumb, P. (1998). *A microfocus X-ray tube used with focusing collimators. J. Appl. Cryst.* **31**, 936–944.

Arndt, U. W. & Stubbings, S. J. (1988). *Miniature ionisation chambers. J. Appl. Cryst.* **21**, 577.

Bailey, R. L. (1978). *The design and operation of magnetic liquid shaft seals.* In *Thermomechanics of Magnetic Fluids*, edited by B. Berkovsky. London: Hemisphere.

Beuville, E., Beche, J.-F., Cork, C., Douence, V., Earnest, J., Millaud, D., Nygren, H., Padmore, B., Turko, G., Zizka, G., Datte, P. & Xuong Ng, H. (1997). *Two-dimensional pixel array sensor for protein crystallography. Proc. SPIE*, **2859**, 85–92.

Bilderback, D. H., Thiel, D. J., Pahl, R. & Brister, K. E. (1994). *X-ray applications with glass-capillary optics. J. Synchrotron Rad.* **1**, 37–42.

Bly, P. & Gibson, D. (1996). *Polycapillary optics focus and collimate X-rays. Laser Focus World*, March issue.

Buras, B. & Tazzari, S. (1984). Editors. *European Synchrotron Radiation Facility.* Geneva: ESRP.

Elliott, A. (1965). *The use of toroidal reflecting surfaces in X-ray diffraction cameras. J. Sci. Instrum.* **42**, 312–316.

Forsyth, J. M. & Frankel, R. D. (1984). *Experimental facility for nanosecond time-resolved low-angle X-ray diffraction experiments using a laser-produced plasma source. Rev. Sci. Instrum.* **55**, 1235–1242.

Fourme, R., Ducruix, A., Ries-Kautt, M. & Capelle, B. (1995). *The perfection of protein crystals probed by direct recording of Bragg reflection profiles with a quasi-planar X-ray wave. J. Synchrotron Rad.* **2**, 136–142.

Franks, A. (1995). *An optically focusing X-ray diffraction camera. Proc. Phys. Soc. London Sect. B*, **68**, 1054–1069.

Genz, H., Graf, H.-D., Hoffmann, P., Lotz, W., Nething, U., Richter, A., Kohl, H., Weickenmeyer, A., Knüpfer, W. & Sellschop, J. P. F. (1990). *High intensity electron channeling and perspectives for a bright tunable X-ray source. Appl. Phys. Lett.* **57**, 2956–2958.

Green, M. (1963). *The target absorption correction in X-ray microanalysis.* In *X-ray Optics and X-ray Microanalysis*, edited by H. H. Pattee, V. E. Cosslett & A. Engström, pp. 361–377. New York and London: Academic Press.

Green, M. & Cosslett, V. E. (1968). *Measurements of K, L and M shell X-ray production efficiencies. Br. J. Appl. Phys. Ser.* 2, **1**, 425–436.

Hofmann, A. (1978). *Quasi-monochromatic synchrotron radiation from undulators. Nucl. Instrum. Methods*, **152**, 17–21.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

Ishimura, T., Shiraiwa, Y. & Sawada, M. (1957). *The input power limit of the cylindrical rotating anode of an X-ray tube. J. Phys. Soc. Jpn*, **12**, 1064–1070.

Kirkpatrick, P. & Baez, A. V. (1948). *J. Opt. Soc. Am.* **56**, 1–13.

Kirz, J. (1974). *Phase zone plates for X-rays and the extreme UV. J. Opt. Soc. Am.* **64**, 301–309.

Kleffer, J. C., Chaker, M., Matte, J. P., Pépin, H., Côté, C. Y., Beaudouin, Y., Johnston, T. W., Chien, C. Y., Coe, S., Mourou, G. & Peyrusse, O. (1993). *Ultra-fast X-ray sources. Phys. Fluids*, **B5**, 2676–2681.

Kohra, K., Ando, M., Natsushita, T. & Hashizume, H. (1978). *Nucl. Instrum. Methods*, **152**, 161–166.

Kumakhov, M. A. & Komarov, F. K. (1990). *Phys. Rep.* **191**, 289–350.

Lemonnier, M., Fourme, R., Rousseaux, F. & Kahn, R. (1978). *X-ray curved-crystal monochromator system at the storage ring DCI. Nucl. Instrum. Methods*, **152**, 173–177.

MacDonald, C. A., Owens, S. M. & Gibson, W. M. (1999). *Polycapillary X-ray optics for microdiffraction. J. Appl. Cryst.* **32**, 160–167.

Milch, J. R. (1983). *A focusing X-ray camera for recording low-angle diffraction from small specimens. J. Appl. Cryst.* **16**, 198–203.

Montel, M. (1957). *X-ray microscopy with catamegonic roof mirrors.* In *X-ray Microscopy and Microradiography*, edited by V. E. Cosslett, A. Engstrom & H. H. Pattee Jr, pp. 177–185. New York: Academic Press.

Müller, A. (1929). *A spinning target X-ray generator and its input limit. Proc. R. Soc. London Ser. A*, **125**, 507–516.

Müller, A. (1931). *Further estimates of the input limits of X-ray generators. Proc. R. Soc. London Ser. A*, **132**, 646–649.

Nagel, D. J. (1980). *Comparison of X-ray sources. Ann. N. Y. Acad. Sci.* **342**, 235–247.

Nave, C., Clark, G., Gonzalez, A., McSweeney, S., Hart, M. & Cummings, S. (1995). *Tests of an asymmetric monochromator to provide increased flux on a synchrotron radiation beam line. J. Synchrotron Rad.* **2**, 292–295.

Oosterkamp, W. J. (1948*a*). *The heat dissipation in the anode of an X-ray tube. Philips Res. Rep.* **3**, 49–59.

Oosterkamp, W. J. (1948*b*). *The heat dissipation in the anode of an X-ray tube. Philips Res. Rep.* **3**, 161–173.

Oosterkamp, W. J. (1948*c*). *The heat dissipation in the anode of an X-ray tube. Philips Res. Rep.* **3**, 303–317.

Osmic Inc. (1998). Sales literature. Osmic Inc., Troy, Michigan, USA.

Padmore, H. A., Ackermann, G., Celestre, R., Chang, C. H., Franck, K., Howells, M., Hussain, Z., Irick, S., Locklin, S., MacDowell, A. A., Patel, J. R., Rah, S. Y., Renner, T. R. & Sandler, R. (1997). *Submicron white-beam focusing using elliptically bent mirrors. Synchrotron Radiat. News*, **10**, 18–26.

Phillips, W. C. (1985). *X-ray sources. Methods Enzymol.* **114**, 300–316.

Piestrup, M. A., Boyers, D. G., Pincus, C. I., Harris, J. L., Maruyama, X. K., Bergstrom, J. C., Caplan, H. S., Silzer, R. M. & Skopik, D. M. (1991). *Quasimonochromatic X-ray source using photoabsorption-edge transition radiation. Phys. Rev. A*, **43**, 3653–3661.

Quintana, J. P. & Hart, M. (1995). *Adaptive silicon monochromators for high-power wigglers; design, finite-element analysis and laboratory tests. J. Synchrotron Rad.* **2**, 119–123.

Schuster, M. & Göbel, H. (1997). *Application of graded multi-layer optics in X-ray diffraction. Adv. X-ray Anal.* **39**, 57–71.

Schwinger, J. (1949). *On the classical radiation of accelerated electrons. Phys. Rev.* **75**, 1912–1925.

Silfhout, R. G. van (1998). *A new water-cooled monochromator at DORIS III. Synchrotron Radiat. News*, **11**, 11–13.

Smither, R. K. (1982). *New methods for focusing X-rays and gamma rays. Rev. Sci. Instrum.* **53**, 131–141.

Winick, H. (1980). *Properties of synchrotron radiation.* In *Synchrotron Radiation Research*, edited by H. Winick & S. Doniach. New York: Plenum.

Yoshimatsu, M. & Kozaki, S. (1977). *High brilliance X-ray source.* In *X-ray Optics*, edited by H.-J. Queisser, ch. 2. Berlin: Springer.

International Tables for Crystallography (2012). Vol. F, Chapter 6.2, pp. 168–176.

Chapter 6.2. Neutron sources

B. P. Schoenborn and R. Knott

6.2.1. Reactors

The generation of neutrons by steady-state nuclear reactors is a well established technique (Bacon, 1962; Kostorz, 1979; Pynn, 1984; Carpenter & Yelon, 1986; Windsor, 1986; West, 1989). Reactor sources that play a major role in neutron-beam applications have a maximum unperturbed thermal neutron flux, $\varphi_{\rm th}$, within the range $1 < \varphi_{\rm th} < 20 \times 10^{14}$ cm^{-2} s^{-1}. A research reactor is essentially a matrix of fuel, coolant, moderator and reflector in a well defined geometry (Fig. 6.2.1.1). The fuel is uranium, and neutron-induced fission in the isotope ^{235}U produces a number of prompt and delayed neutrons (slightly more than a total of two) and, on average, one of these is required to maintain the steady-state chain reaction (*i.e.* criticality). The heat generated (~200 MeV per fission event) must be removed, hence the need for an efficient coolant. In practice, the simultaneous requirements of complex thermal hydraulics and nuclear reaction kinetics must be addressed. The neutrons produced in the fission event have a mean energy of ~1 MeV, and a material is required to reduce this energy to ~25 MeV to take advantage of the larger fission cross section of ^{235}U in the 'thermal' energy range. Such a moderator is composed of a material rich in light nuclei, so that a large fraction of the neutron energy is transferred per collision.

There is an inherent maximum in neutron flux density imposed by the fission process (the number of excess neutrons produced per fission event), by the reduced density of neutron-generation material required for cooling purposes and by the heat-removal capacity of suitable coolants. Detailed design of reactor systems is essential to obtain the correct balance.

6.2.1.1. Basic reactor physics

Reactor physics is the theoretical and experimental study of the neutron distributions in the energy, spatial and time domains (Soodak, 1962; Jakeman, 1966; Akcasu *et al.*, 1971; Glasstone & Sesonske, 1994). The fundamental relation describing neutron kinetics is the Boltzmann transport equation (*e.g.* Spanier & Gelbard, 1969; Stamm'ler & Abbate, 1983; Weisman, 1983; Lewis & Miller, 1993). In theory, the transport equation describes the life of the neutron from its birth as a high-energy component of the fission process, through the various diffusion and moderation processes, until its ultimate end in (i) the chain reaction, (ii) leakage into beam tubes, or (iii) parasitic absorption (Williams, 1966). In practice, the complex nuclear reactions and the geometrical configurations of the component materials are such that a rigorous theoretical analysis is not always possible, and simplifying approximations are necessary. Nevertheless, well proven algorithms have been developed, and many have been included in computer codes (*e.g.* Hallsall, 1995).

On examination of the factors that influence the neutron flux distribution, there are three distinct but interdependent functions performed by the coolant, moderator and reflector. The coolant/moderator is of major importance in the fuelled region to sustain optimum conditions for the chain reaction, and the moderator/reflector is important in the regions surrounding the central core (Section 6.2.1.2). It should be noted that reactors for neutron-beam applications must be substantially undermoderated in order to provide a fast neutron flux at the edge of the core, which can be thermalized at the entry to the beam tubes. Most research reactors use H_2O or D_2O as the coolant/moderator.

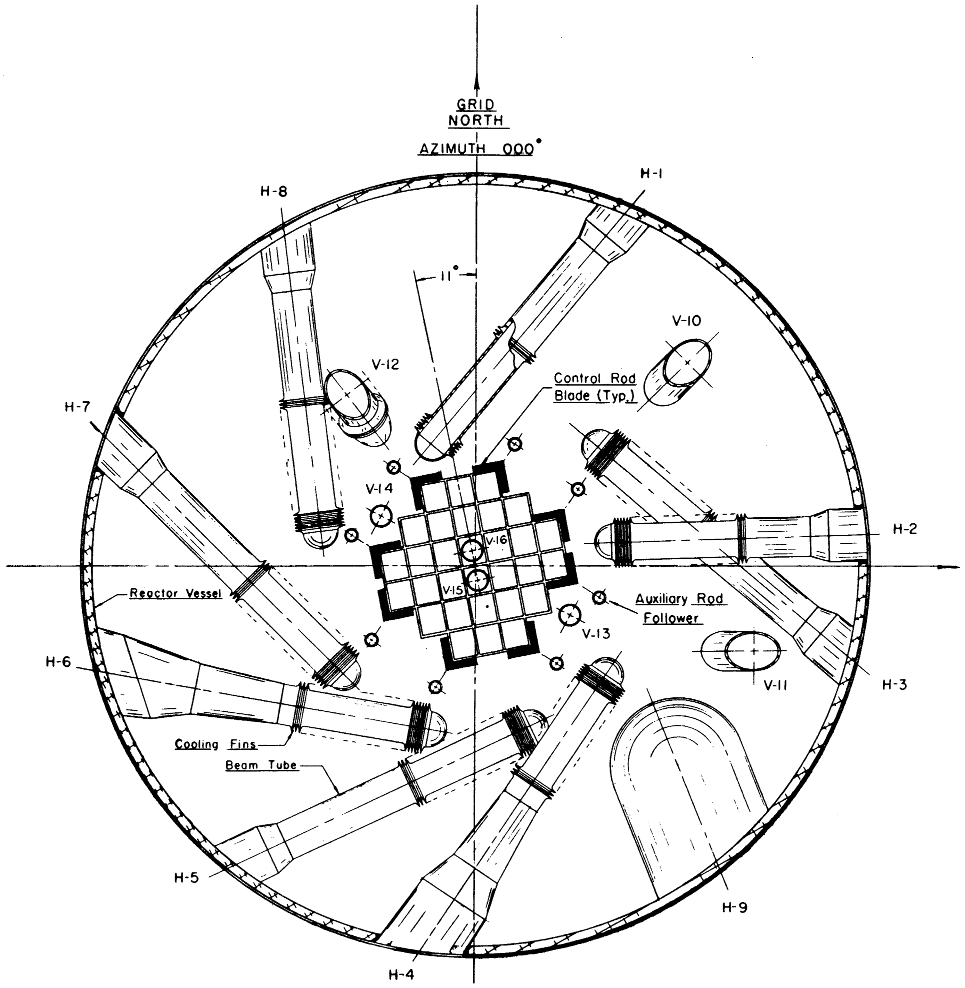


Figure 6.2.1.1
Schematic of the High Flux Beam Reactor at the Brookhaven National Laboratory (USA). The central core region (48 cm in diameter and 58 cm high) contains 28 fuel elements in an array surrounded by an extended D_2O moderator/reflector region. The diameter of the reactor vessel is approximately 2 m. All but one of the beam tubes are tangentially oriented with respect to the core. The cold neutron source is located in the H9 beam tube.

Copyright © 2012 International Union of Crystallography

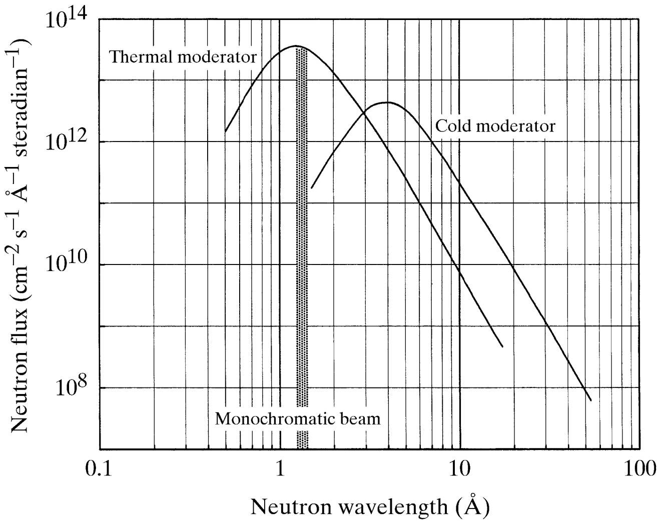


Figure 6.2.1.2

Neutron wavelength distributions for a thermal (310 K) and a cold (30 K) neutron moderator in a 'typical' dedicated beam reactor. The Maxwellian distribution merges with a $1/E$ slowing-down distribution at shorter wavelengths. A wavelength distribution for a monochromatic beam application on a thermal source is illustrated. It should be noted that, depending on the value of the mean wavelength for the monochromatic beam, harmonic contamination may be significant.

6.2.1.2. Moderators for neutron scattering

The moderator/reflector serves to modify the energy distribution of fast neutrons leaking from the central core, returning a significant number of thermalized neutrons to the core region to provide for criticality with a smaller inventory of fuel and providing excess neutrons for a range of applications, including neutron-beam applications. The moderator/reflector may be H_2O, D_2O, Be, graphite or a combination of these. Almost all choices have been used; however, the optimum is not achieved with any choice, and priorities must be set in terms of neutron-beam performance, other source activities and the reactor fuel cycle.

For neutron-beam applications, H_2O is the preferred reflector material since the moderation length is large and the absorption cross section low. Consequently, the thermal neutron flux peak is rather broad and occurs relatively distant from the core. While the exact position of the peak depends on the reactor core design, the peak width has a major impact on beam-tube orientation (Section 6.2.1.3). The combination of D_2O coolant/moderator–D_2O moderator/reflector provides a distinct advantage; however, for a number of technical reasons, the H_2O–D_2O combination is becoming more common, with the D_2O in a closed vessel surrounding the central H_2O-cooled core.

6.2.1.2.1. Thermal moderators

Neutrons thermalized within a 'semi-infinite' moderator/reflector typical of a steady-state reactor source establish an equilibrium Maxwellian energy distribution characterized by the temperature (T) of the moderator (Fig. 6.2.1.2). The wavelength, λ_m, at which the above distribution has a maximum is given by

$$\lambda_m = h/(5k_BTm_n)^{1/2}.$$

Depending on the width of the moderator and its composition, the Maxwellian distribution merges with the $1/E$ slowing-down distribution from the reactor core to give a total distribution at the beam-tube entry.

Clearly, the neutron wavelength distribution will depend on the local equilibrium conditions. Since steady-state reactors typically operate with moderator/reflector temperatures in the range 308–323 K, the corresponding λ_m is about 1.4 Å (Fig. 6.2.1.2). However, it is possible to alter the neutron distribution by re-thermalizing the neutrons in special moderator regions, which are either cooled significantly below or heated significantly above the average moderator temperature. One such device of prime importance is the cold moderator in the form of a cold source.

6.2.1.2.2. Cold moderators

The thermal neutron distribution shown in Fig. 6.2.1.2 is not ideal for all experiments, since the flux of 5 Å neutrons is almost two orders of magnitude less than the peak. The solution is to introduce a cold region in the moderator/reflector. This is typically a volume of liquid H_2 or D_2 at ~10–30 K, and a Maxwellian distribution around this temperature results in a λ_m of ~4 Å. The geometric design of a cold-source vessel has been shown to be very important, with substantial gains in neutron flux obtained by innovative design. A re-entrant geometry approximately 20 cm in diameter filled with liquid D_2 at 10 K provides optimum neutron thermalization with superior coupling to neutron guides (Section 6.2.1.3.5) (Ageron, 1989; Lillie & Alsmiller, 1990; Alsmiller & Lillie, 1992).

6.2.1.3. Beamline components

Until recently, instrument design has been largely based on experience; however, in many cases, it is now possible to formulate a comprehensive description of the instrument and explore the impact of various parameters on instrument performance using an extensive array of computational methods (Johnson & Stephanou, 1978; Sivia *et al.*, 1990; Hjelm, 1996). In practice, it is the instrument design that provides access to the fundamental scattering processes, as briefly outlined in the following.

If a neutron specified by a wavevector $\mathbf{k}_1$ is incident on a sample with a scattering function $S(\mathbf{Q}, \omega)$, all neutron scattering can be reduced to the simple form

$$\frac{d^2\sigma}{d\Omega\, dE} = AS(\mathbf{Q}, \omega),$$

where A is a constant containing experimental information, including instrumental resolution effects. The basic quantity to be measured is the partial differential cross section, which gives the fraction of neutrons of incident energy E scattered into an element of solid angle Ω with an energy between E' and $E' + dE'$. The momentum transfer, $\mathbf{Q}$, is given in Fig. 6.2.1.3(*a*). The primary aim of a neutron-scattering experiment is to measure $\mathbf{k}_2$ to a predetermined precision (Bacon, 1962; Sears, 1989) (Fig. 6.2.1.3*b*). A generic neutron-scattering instrument used to achieve this aim is illustrated in Fig. 6.2.1.4. The instrument resolution function will be determined by uncertainties in $\mathbf{k}_1$ and $\mathbf{k}_2$, which are a direct consequence of (i) measures to increase the neutron flux at the sample position (to maximize wavelength spread, beam divergence, monochromator mosaic, for example) and (ii) uncertainties in geometric parameters (flight-path lengths, detector volume *etc.*).

6.2.1.3.1. Collimators and filters

In general, neutron-beam applications are flux-limited, and a major advantage will be realized by adopting advanced techni-

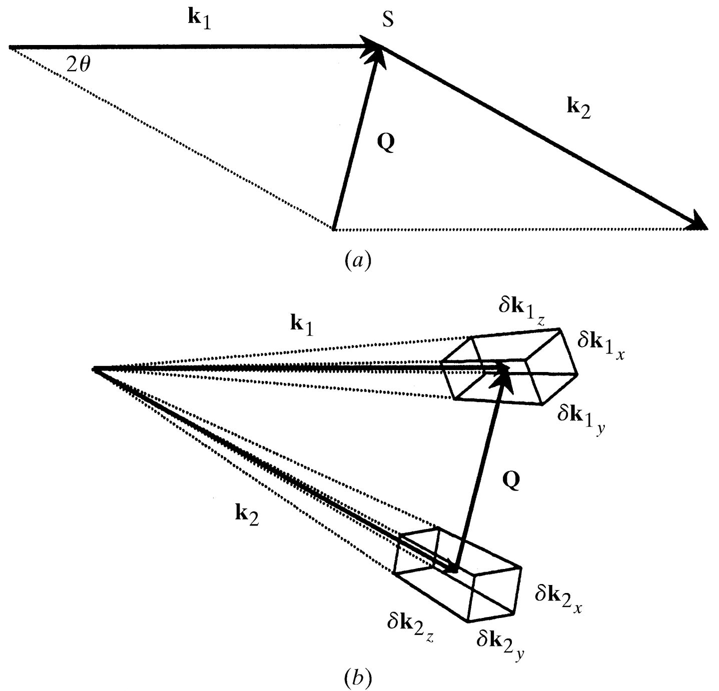


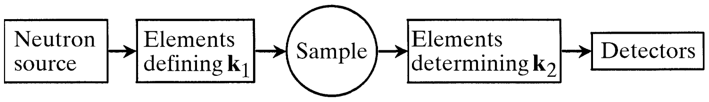
Figure 6.2.1.3

(*a*) Schematic vector diagram for an elastic neutron-scattering event. A neutron, $\mathbf{k}_1$, is incident on a sample, S, and a scattered neutron, $\mathbf{k}_2$, is observed at an angle 2θ leading to a momentum transfer, **Q**. (*b*) Schematic of an elastic neutron-scattering event illustrating the consequences of uncertainty in defining the incident neutron, $\mathbf{k}_1$, and determining the scattered neutron, $\mathbf{k}_2$. The volumes ($\delta\mathbf{k}_{1_x}$, $\delta\mathbf{k}_{1_y}$, $\delta\mathbf{k}_{1_z}$) and ($\delta\mathbf{k}_{2_x}$, $\delta\mathbf{k}_{2_y}$, $\delta\mathbf{k}_{2_z}$) constitute the instrument resolution function.

ques in flux utilization. A reactor neutron source is large, and stochastic processes dominate the generation, moderation and general transport mechanisms. Because of radiation shielding, background reduction and space requirements for scattering instruments, reactor beam tubes have a minimum length of 3–4 m and cannot exceed a diameter of about 30 cm. The useful flux at the beam-tube exit is thus between 10^{-5} and 10^{-6} times the isotropic flux at its entry.

The most important aspect of beam-tube design, other than the size, is the position and orientation of the tube with respect to the reactor core. The widespread utilization of D_2O reflectors enables significant gains to be obtained from tangential tubes with maximal thermal neutron flux, and minimal fast neutron and γ fluxes. A similar result is accomplished in a split-core design by orienting the beam tubes toward the unfuelled region of the reactor core (Prask *et al.*, 1993).

There are two major groups of neutron-scattering instruments: those located close to and those located some distance from the neutron source. The first group is located to minimize the impact of the inverse square law on the neutron flux, and the second group is located to reduce background and provide more instrument space. In both cases, the first beamline component is a collimator to extract a neutron beam of divergence α from the reactor environment. The collimator is usually a beam tube of suitable dimensions for fully illuminating the wavelength-selection device. The angular acceptance of a collimator is determined strictly by the line-of-sight geometry between the source and the monochromator. Some geometric focusing may be appropriate, and a Soller collimator may be used to reduce α without reducing the beam dimensions. For technical reasons, the primary collimator is essentially fixed in dimensions and secondary collimators of adjustable dimensions may be required in more accessible regions outside the reactor shielding.



Neutron source
Elements defining $\mathbf{k}_1$
Sample
Elements determining $\mathbf{k}_2$
Detectors

Figure 6.2.1.4

A generic neutron-scattering instrument illustrating the classes of facilities and operators important to instrument design and assessment. Each class should be optimized and integrated into the overall instrument description.

Neutron-beam filters are required for two main reasons: (i) to reduce beam contamination by fast neutron and γ radiation and (ii) to reduce higher- or lower-order harmonics from a monochromatic beam. Numerous single-crystal, polycrystalline and multilayer materials with suitable characteristics for filter applications are available (*e.g.* Freund & Dolling, 1995).

6.2.1.3.2. *Crystal monochromators*

The equilibrium neutron-wavelength distribution (Fig. 6.2.1.2) is a broad continuous distribution, and in most experiments it is necessary to select a narrow band in order to define $\mathbf{k}_1$. Neutron-wavelength selection can be achieved by Bragg scattering using single crystals to give well defined wavelengths; by polycrystalline material to remove a range of wavelengths; by a mechanical velocity selector; or by time-of-flight methods. The method chosen will depend on experimental requirements for the wavelength, λ, and wavelength spread, $\Delta\lambda/\lambda$.

Neutrons incident on a perfect single crystal of given interplanar spacing d will be diffracted to give specific wavelengths at angle 2θ according to the Bragg relation. A neutron beam, α_1, incident on a single crystal of mosaicity β will provide

$$\Delta\lambda/\lambda = [(\cot\theta\cdot\theta)^2 + (\Delta d/d)^2]^{1/2},$$

where

$$(\Delta\theta)^2 = \frac{\alpha_1^2\alpha_2^2 + \beta^2(\alpha_1^2 + \alpha_2^2)}{\alpha_1^2 + \alpha_2^2 + 4\beta^2}$$

and α_2 is the divergence of the (unfocused) diffracted beam.

Crystals for neutron monochromators must not only have a suitable d, but also high reflectivity and adequate β. Under these conditions, neutron beams with a $\Delta\lambda/\lambda$ of a few per cent are obtained. Typical crystals are Ge, Si, Cu and pyrolytic graphite. In order to increase the neutron flux at the sample, a number of mechanisms have been developed. These include focusing monochromatic crystals, frequently using Si or Ge (Riste, 1970; Mikula *et al.*, 1990; Copley, 1991; Magerl & Wagner, 1994; Popovici & Yelon, 1995), as well as stacked composite wafer monochromators (Vogt *et al.*, 1994; Schefer *et al.*, 1996).

One limitation of the use of crystal monochromators is the absence of suitable materials with large d. Indeed, the longest useable λ diffracted from pyrolytic graphite or Si is ~5–6 Å.

6.2.1.3.3. *Multilayer monochromators and supermirrors*

Multilayers are especially useful for preparing a long-wavelength neutron beam from a cold source and for small-angle scattering experiments in which $\Delta\lambda/\lambda$ of about 0.1 is acceptable (Schneider & Schoenborn, 1984). Multilayer monochromators are essentially one-dimensional crystals composed of alternating layers of neutron-different materials (*e.g.* Ni and Ti) deposited on a substrate of low surface roughness. In order to produce multilayers of excellent performance, uniform layers are required with low interface roughness, low interdiffusion between layers and high scattering contrast. Various modifications (*e.g.* carbonation, partial hydrogenation) to the pure Ni and Ti bilayers improve the performance significantly by fine tuning the layer uniformity and contrast (Mâaza *et al.*, 1993). The minimum practical *d*-spacing is ~50 Å and a useful upper limit is ~150 Å. Multilayer monochromators have high neutron reflectivity (>0.95 is achievable), and their angular acceptance and bandwidth can be selected to produce a neutron beam of desired characteristics (Saxena & Schoenborn, 1977, 1988; Ebisawa *et al.*, 1979; Sears, 1983; Schoenborn, 1992*a*).

The supermirror, a development of the multilayer monochromator concept, consists of a precise number of layers with graded *d*-spacing. Such a device enables the simultaneous satisfaction of the Bragg condition for a range of λ and, hence, the transmission of a broader bandwidth (Saxena & Schoenborn, 1988; Hayter & Mook, 1989; Böni, 1997).

Polarizing multilayers and supermirrors (Schärpf & Anderson, 1994) facilitate valuable experimental opportunities, such as nuclear spin contrast variation (Stuhrmann & Nierhaus, 1996) and polarized neutron reflectometry (Majkrzak, 1991; Krueger *et al.*, 1996). Supermirrors consisting of Co and Ti bilayers display high contrast for neutrons with a magnetic moment parallel to the saturation magnetization and very low contrast for the remainder. With suitable modification of the substrate to absorb the antiparallel neutrons, a polarizing supermirror will produce a polarized neutron beam (polarization >90%) by reflection.

6.2.1.3.4. *Velocity selectors*

The relatively low speed of longer-wavelength neutrons (~600 m s^{-1} at 6 Å) enables wavelength selection by mechanical means (Lowde, 1960). In general, there are two classes of mechanical velocity selectors (Clark *et al.*, 1966). Rotating a group of short, parallel, curved collimators about an axis perpendicular to the beam direction will produce a pulsed neutron beam with λ and $\Delta\lambda/\lambda$ determined by the speed of rotation. This is a Fermi chopper. An alternate method is to translate short, parallel, curved collimators rapidly across the neutron beam, permitting only neutrons with the correct trajectory to be transmitted. This is achieved in the helical velocity selector, where the neutron wavelength is selected by the speed of rotation and $\Delta\lambda/\lambda$ can be modified by changing the angle between the neutron beam and the axis of rotation (Komura *et al.*, 1983). The neutron beam is essentially continuous, the resolution function is approximately triangular and the overall neutron transmission efficiency exceeds 75% in modern designs (Wagner *et al.*, 1992).

6.2.1.3.5. *Neutron guides*

In order for a collimator to be effective, its walls must absorb all incident neutrons. The angular acceptance is strictly determined by the line-of-sight geometry. Neutron guides can be used to improve this acceptance dramatically and to transport neutrons with a given angular distribution, almost without intensity loss, to regions distant from the source (Maier-Leibnitz & Springer, 1963). The basic principle of a guide is total internal reflection. This occurs for scattering angles less than the critical angle, θ_c, given by

$$\theta_c = 2(1-n)^{1/2},$$

where n is the (neutron) index of refraction related to the coherent scattering length, b, of the wall material, *viz*,

$$n = 1 - (\lambda^2 \rho b / 2\pi),$$

where ρ is the atom number density (in cm^{-3}). Among common materials, Ni with $b = 1.03 \times 10^{-12}$ cm, in combination with suitable physico-chemical properties, provides the best option, with a critical angle $\theta_c = 0.1\lambda$ (in Å). The dependence of θ_c on λ implies that guides are more effective for long-wavelength neutrons. With the introduction of supermirror guides with up to four times the θ_c of bulk Ni, both thermal and cold neutron beams are being transported and focused with high efficiency (Böni, 1997).

While a straight guide transports long wavelengths efficiently, it continues to transport all neutrons within the critical angle, including non-thermal neutrons emitted within the solid angle of the guide. This situation may be modified significantly by introducing a curvature to the guide. Since a curved neutron guide provides a form of spectral tailoring (cutoff or bandpass filters), simulation is a distinct advantage in exploring the impact of guide geometry on neutron-beam quality (van Well *et al.*, 1991; Copley & Mildner, 1992; Mildner & Hammouda, 1992).

6.2.1.4. *Detectors*

The detection of thermal neutrons is a nuclear event involving one of only a few nuclei with a sufficiently large absorption cross section (^{3}He, ^{10}B, ^{6}Li, Gd and ^{235}U). The secondary products (fragments, charged particles or photons) from the primary nuclear event are used to determine the location. Depending on the geometry of the instrument, either a spatially integrating or a position-sensitive detector is required (Convert & Forsyth, 1983; Crawford, 1992; Rausch *et al.*, 1992). The relatively weak neutron source is driving instrument design towards maximizing the number of neutrons collected per unit time, and, in many cases, this leads to the use of multiwire or position-sensitive detectors. The main performance characteristics for detector systems are position resolution, number of resolution elements, efficiency, parallax, maximum count rate, dynamic range, sensitivity to γ background and long-term stability.

6.2.1.4.1. *Multiwire proportional counters*

The principles of a multiwire proportional counter (MWPC) are well established (Sauli, 1977) and have wide application. For thermal neutron detection (Radeka *et al.*, 1996), the reaction of choice is

$$^3\mathrm{He} + \mathrm{n} \rightarrow {}^3\mathrm{H} + \mathrm{p} + 764\ \mathrm{keV}.$$

The 191 keV triton and the 573 keV proton are emitted in opposite directions and create a charge cloud whose dimensions are determined primarily by the pressure of a stopping gas. Depending on the work function of the gas mixture, approximately 3×10^4 electron–ion pairs are created. Low-noise gas amplification of this charge cloud occurs in an intense electric field created in the vicinity of the small diameter (20–30 μm) anode wires (Radeka, 1988). Typical gas gains of ~10–50 lead to a total charge on the anode of ~50–100 fC. The efficiency of the detector is determined by the pressure of ^{3}He, and the spatial resolution and count-rate capability are determined by the detector geometry and readout system. The event decoding is selected from the time difference (Borkowski & Kopp, 1975), charge division (Alberi *et al.*, 1975), centroid-finding filter (Radeka & Boie, 1980), or wire-by-wire techniques (Jacobé *et al.*, 1983; Knott *et al.*, 1997). Present MWPC technology offers opportunities and challenges to design a detector system that is totally integrated into the instrument design and optimizes data collection rate and accuracy (Schoenborn *et al.*, 1985, 1986; Schoenborn, 1992*b*).

A concept related to the MWPC is the micro-strip gas chamber (MSGC). With the MSGC, the general principles of gas detection and amplification apply; however, the anode is deposited on a suitable substrate (Oed, 1988, 1995; Velletaz *et al.*, 1997). The MSGC can potentially improve the performance of the MWPC in some applications, particularly with respect to spatial resolution and count-rate capability.

6.2.1.4.2. *Image plates*

The principles underlying the operation of an image plate (IP) are presented in detail in Chapter 7.2. Briefly, the important difference between an IP for X-ray and neutron detection is the presence of a converter (either Gd_2O_3 or ^6Li). The role of the converter is to capture an incoming neutron and create an event within the IP that mimics the detection of an X-ray photon. For example, neutron capture in Gd produces conversion electrons that exit the Gd_2O_3 grains, enter neighbouring photostimulated luminescence (PSL) material and create colour centres to form a latent image (Niimura *et al.*, 1994; Takahashi *et al.*, 1996). A neutron IP may have a virtually unlimited area and a shape limited only by the requirement to locate the detection event in a suitable coordinate system. With a neutron-detection efficiency of up to 80% at ~1–2 Å, a dynamic quantum efficiency of ~25–30% can be obtained. The dynamic range is intrinsically $1{:}10^5$. The spatial resolution is primarily limited by scattering processes of the readout laser beam, and measured line spread functions are typically 150–200 µm. The γ sensitivity is high and may restrict the application to instruments with low ambient γ background.

Neutron IPs are integrating devices well suited to data-acquisition techniques with long accumulation times, such as Laue diffraction (Niimura *et al.*, 1997) and small-angle scattering. On-line readout is a distinct advantage (Cipriani *et al.*, 1997).

6.2.1.5. *Instrument resolution functions*

For accurate data collection, the instrument smearing contribution to the data must be known with some certainty, particularly when data are collected over an extended range with multiple instrument settings. A balance must be struck between instrument smearing and neutron flux at the sample position; however, careful instrument design can produce: (i) a good signal-to-background ratio, thereby partially offsetting the flux limitation, and (ii) facilities and procedures for determining the instrument resolution function (Johnson, 1986).

As an example, instrumental resolution effects in the small-angle neutron scattering (SANS) technique have been investigated in some detail. A 'typical' SANS instrument is located on a cold neutron source with an extended (and often variable) collimation system. The sample is as large as possible and the detector is large with low spatial resolution. The instrument is best described by pin-hole geometry. Three major contributions to the smearing of an ideal curve are: (i) the finite λ, (ii) $\Delta\lambda/\lambda$ of the beam and (iii) the finite resolution of the detector. Indirect Fourier transform, Monte Carlo and analytical methods have been developed to analyse experimental data and predict the performance of a given combination of resolution-dependent elements (*e.g.* Wignall *et al.*, 1988; Pedersen *et al.*, 1990; Harris *et al.*, 1995).

6.2.2. Spallation neutron sources

Another phenomenon, quite different from the fission process (Section 6.2.1), that will produce neutrons uses high-energy particles to interact with elements of medium to high mass numbers. This process, called spallation, was first demonstrated by Seaborg and Perlman, who showed that the bombardment of nuclei by high-energy particles results in the emission of various nucleons. The nuclear processes involved in spallation (Prael, 1994) are complex and are summarized in Fig. 6.2.2.1. These processes have been investigated in some detail, and excellent background information is available (Hughes, 1988; Carpenter, 1977; Windsor, 1981). Present-day spallation sources typically use high-energy protons from an accelerator to bombard a heavy-metal target, such as W or U, and come in two types, using either a pulsed proton beam (*e.g.* ISIS or LANSCE) or a 'continuous' proton beam (SINQ).

The high-energy neutrons produced by spallation are moderated in a reflector region to intermediate energies and then reduced to thermal energies in a hydrogenous medium called the moderator (Russell *et al.*, 1996). These thermal neutrons are then extracted *via* beam pipes. A typical layout of a target system with reflectors and moderators is shown in Fig. 6.2.2.2.

The neutrons produced by the proton pulse travel along beam pipes as a function of their velocity, proportional to their energy.

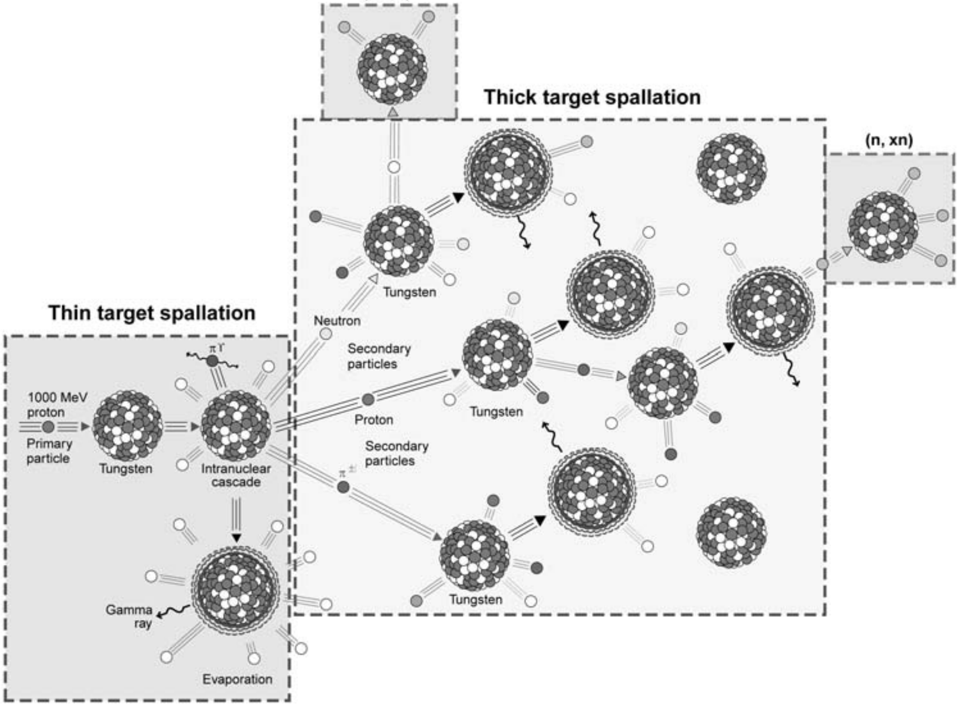


Figure 6.2.2.1
Schematic presentation of the various nuclear processes encountered in spallation. The numerical analysis of these processes is carried out by two Monte Carlo-based codes – the *LAHET* code models the higher-energy nuclear interactions, while the *HMCNP* code models the thermal interactions and the transport of neutrons to the sample.

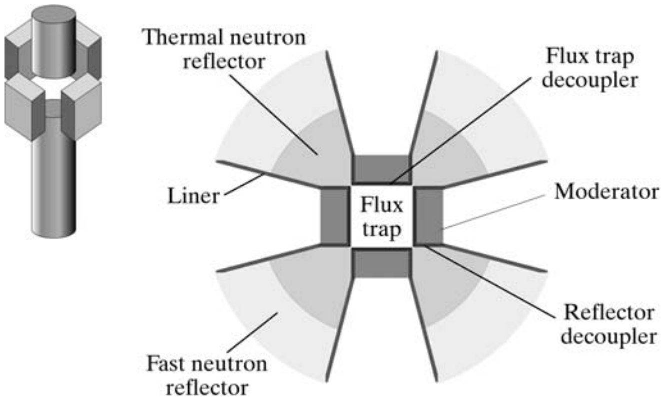


Figure 6.2.2.2
Schematic diagram of a spallation target module depicting the inner Be and outer Pb reflector. Moderators are positioned close to the flux trap, which is between the upper and lower tungsten targets. This schematic includes the location of the various decoupling agents (Cd) for fully decoupled moderators. For the partially coupled moderator system being fabricated for a protein-crystallography station at Los Alamos, the depicted decouplers affecting the given moderator are removed and replaced by a single decoupling layer placed between the outer Pb and the inner Be reflector. This can be done with a split-target arrangement that utilizes a two-tiered moderator system, permitting coupled moderators to be on one tier and decoupled moderators on the other.

At a given distance from the target, neutrons of different energies are observed to arrive as a function of time, with the short-wavelength neutrons arriving first, followed by the longer-wavelength neutrons. Diffraction experiments are therefore carried out with pulsed 'polychromatic' neutron beams as a function of time; a time-resolved Laue pattern results. Clearly, the energy resolution of these beams depends on the volume of the neutron source. To achieve high energy resolution, the volume from which thermal neutrons are extracted is limited to the moderator by suitable use of liners and poisons (Fig. 6.2.2.2) that prevent thermal neutrons produced in the reflector from streaming into the beam pipe (decoupled moderators). The use of liners and poisons achieves high energy resolution, but at the expense of flux. The omission or reconfiguration of liners and poisons allows higher flux, but results in lower energy (wavelength) resolution (see Section 6.2.2.2).

6.2.2.1. *Spallation neutron production*

Pulsed spallation neutrons are produced by protons generated by a particle accelerator (linac) with a frequency typically in the range 10 to 120 Hz. The proton pulses are often shaped in compressor rings to shorten the pulses from the millisecond range to less than a microsecond in duration, with currents reaching the sub-milliampere range at energies of 800 MeV or higher. The planned new spallation source at the Oak Ridge National Laboratory will have a proton energy of 1.2 GeV with a power of 1 MW and a repetition frequency of 60 Hz. The high-energy protons hit a cooled target (typically W or U) and produce high- and medium-energy neutrons in equivalent bursts. The fast neutrons are moderated in the surrounding reflector and returned to the moderator (Fig. 6.2.2.2). In addition to direct collision interactions, the high-energy neutrons also produce low-energy neutrons *via* (n, *x*n) reactions in the Pb and Be reflectors. Final moderation to the thermal energies used for diffraction experiments is completed *via* interactions with light elements, such as H_2O or liquid H_2, in the moderator module.

6.2.2.2. *Moderators*

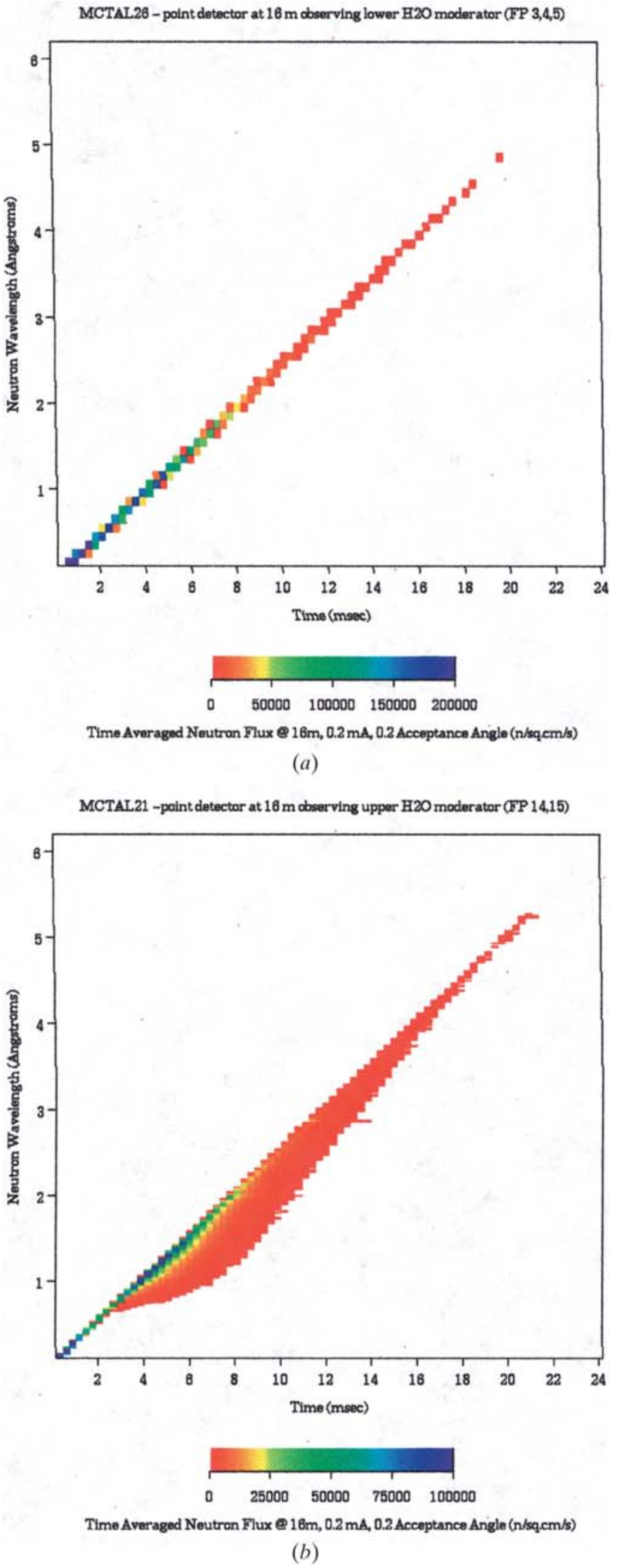
Moderators for pulsed spallation neutron sources are nearly always composed of hydrogenous material of about 1 l in volume. Either a thermal or fast reflector surrounds the moderator. Reflectors composed of materials with strong neutron slowing-down properties, such as Be or D_2O, are called thermal reflectors; fast reflectors are composed of materials with weaker slowing-down powers, such as Pb or Ni. In order to retain a narrow pulse width in time, thermal neutrons produced in the reflector region are prevented from reaching the moderator module by judicious use of liners and poisons (typically Cd or Gd) that allow transmission of fast (high and intermediate energy) neutrons, but are opaque to thermal neutrons. Such a moderator arrangement is said to be decoupled, and all thermal neutrons extracted by the beam pipe originate in the moderator itself. For a 0.25 µs-long proton pulse, the target (W or U) produces a fast neutron burst of about 0.5 µs in duration. These very high energy neutrons are slowed down in the reflector and are reflected back into the moderator to produce a thermal neutron pulse of about 1 µs duration. Since thermal neutrons produced in the reflector are prevented from reaching the moderator by the use of liners and poisons, the experiment sees only thermal neutrons originating in the moderator.

By moving the decoupling and poison layers away from the moderator and into the reflector, one can redefine how actively a reflector communicates *via* neutrons with a moderator and how some of the thermal neutrons produced in the reflector are extracted (Schoenborn *et al.*, 1999). The result is an increase in flux, both peak- and time-integrated, but at the expense of the sharply defined time distribution. With the elimination of all liners and poisons, a fully coupled system is obtained with a flux gain of about 6× but with poorer wavelength resolution. The wavelength distribution at a given distance from the moderator is shown in Fig. 6.2.2.3(*a*) for the fully decoupled case and in Fig. 6.2.2.3(*b*) for the fully coupled case. For a decoupled moderator, the slowing-down power of the reflector is not as critical as it is for the coupled one. In the coupled moderator, it is beneficial to use a thermal reflector in the volume immediately surrounding the moderator because this enhances the peak thermal neutron flux. The decay constant of the neutron pulse can be tailored to match the diffractometer resolution by using a composite reflector composed of an inner thermal reflector and an outer fast reflector. The outer reflector can have a moderate thermal neutron absorption cross section, or the inner reflector can be decoupled from the outer reflector in the same manner that a moderator is decoupled from a reflector. The decay constant can then be varied by simply adjusting the size of the inner reflector (Russell *et al.*, 1996). The wavelength or energy distribution of thermal neutrons produced in the moderator is dependent on the temperature of the moderating medium, as described in Section 6.2.1.2.

For neutron protein crystallography, a moderator with an intermediate temperature between a cold and thermal moderator would be most appropriate. This can be achieved with a composite moderator composed of a thermal and a cold moderator in a symbiotic configuration, or a cold methane system.

6.2.2.3. *Beamline optics*

A chosen wavelength band (say, 1 to 6 Å) is selected by the use of rotating disks (called choppers) composed of neutron absorbing and transparent material. These choppers are synchronized to the proton pulse. The T0 chopper can open the



(*a*)

(*b*)

Figure 6.2.2.3
Neutron flux given as a time–wavelength spectrum for (*a*) a fully decoupled system and (*b*) for a fully coupled system. Both spectra are based on Monte Carlo codes (*LAHET* and *HMCNP*) and are calculated for a target-to-sample distance of 10 m. Comparison of such Monte Carlo results calculated using the geometry of an existing beamline shows agreement with measured values to within 10%.

beam a short time after the impact of the proton pulse and stop the high- and intermediate-energy radiation from reaching the sample and the detector. The T1 chopper can select the long-wavelength edge and prevent frame overlap. Since the T0 chopper is designed to stop the initial γ and high-energy neutron radiation, it is usually made of thick (30 cm) blades of Ni, while the T1 chopper is simpler in construction since it is designed to stop only thermal neutrons.

The flight-path lengths of relevant spallation neutron instruments are quite long; the Los Alamos Spallation Neutron Source has a 28 m path length for its protein crystallography station on a partially decoupled moderator. For a fully decoupled system, a flight-path length of 10 m would provide adequate energy resolution.

For protein crystallography, a beam divergence matched to the mosaicity of the crystal provides the best peak-to-background ratio. For such cases, a beam divergence of $\pm 0.1^\circ$ can be achieved using circular collimating disks of Boral or Boron-Poly to form a cone that views most of the moderator (typically 12×12 cm) and channels the neutron beam onto the detector with a final aperture of millimetre dimensions. Another approach uses focusing mirrors, and calculations show that toroidal geometry will produce a gain in intensity of 1.5 to 2 times, depending on flight distance and beam divergence (Schoenborn, 1992*a*).

6.2.2.4. *Time-of-flight techniques*

Because of the time structure inherent at a spallation source, diffraction experiments are carried out as a function of time and use a large part of the neutron energy spectrum. For protein crystallography, this wavelength range might cover from 1 to 5 Å, depending on the unit-cell size and the moderator used. This is particularly advantageous and allows the collection of data in a quasi-Laue fashion (Schoenborn, 1992*a*) without the drawback of spot overlap normally encountered in Laue patterns. Data are collected in a stroboscopic fashion, synchronized to the pulsed nature of the source, with each separately recorded time frame producing a Laue pattern from a narrow, gradually increasing wavelength band. The summation of all time frames will produce a true Laue pattern. The collection and analysis of these quasi-Laue patterns (time frames) will eliminate spot overlap and yield a greatly improved peak-to-background ratio, since the integrated background is produced only by the small wavelength band responsible for a particular diffraction peak.

6.2.2.5. *Data-collection considerations*

Single-event-counting multiwire chambers with centroid-finding electronics (introduced in Section 6.2.1.4) are well suited for the type of time-sliced data collection that is mandatory for spallation neutron instruments. For large, high-resolution, multi-segmented detectors collecting about 100 time slices per cycle, data memories in the order of 100 million pixels are required. The number of time slices that needs to be collected to produce the optimum peak-to-background ratio depends on the characteristics (wavelength bandwidth) of the coupled or decoupled moderator.

Data-integration techniques are similar to those for the classic reactor case (Section 6.2.1.5), but contain a time (wavelength) dimension and no crystal stepping. The crystal is stationary and the reflection is 'scanned' as a function of time by the wavelength band. Time-dependent reflection overlap, caused by long pulse decay (particularly observed in fully coupled moderators), can be a problem. Such overlaps can be minimized by using a partially coupled moderator (Schoenborn *et al.*, 1999).

6.2.3. Summary

In the preceding sections, a brief overview has been presented of (i) the two main types of neutron sources and (ii) some of the primary components required to prepare a neutron beam for a neutron-scattering instrument. It has been assumed that as well as macromolecular crystallography, membrane and fibre diffraction, small-angle neutron scattering (see Chapter 19.4) is of interest. From a structural-biology user perspective, the advantages and disadvantages of reactor-based and spallation-source-based facilities are difficult to assess, since only very limited use of spallation sources has been documented. Direct comparisons between the performances of neutron-scattering instruments and sources are difficult, and would undoubtedly change as facilities are progressively upgraded (Carpenter & Yelon, 1986; Richter & Springer, 1998). Calculations show, however, that the use of time-of-flight techniques with partially coupled moderators on a spallation neutron source is ideal for structural-biology diffraction studies and promises to yield an effective gain of an order of magnitude in intensity (Schoenborn, 1996). When the protein crystallographic diffraction instrument now being built at LANSCE is completed in 2000, a more meaningful comparison will be possible between a premier spallation-source-based instrument and comparable reactor-based instruments.

In summary, the neutron source plays a pivotal role in the design and utility of an experiment in macromolecular crystallography, membrane and fibre diffraction, and small-angle neutron scattering. However, innovative design of the scattering instrument using the latest technology (*e.g.* image plates or large MWPCs) can partially offset certain negative impacts of the source and make an enormous difference to the instrument as a user facility. In general, neutron sources are national or regional facilities and consequently carry special requirements for user access. Therefore, a local, well equipped, medium-flux neutron source may be more suitable to test potential experiments and the premier international facility should be used only where required.

References

Ageron, P. (1989). *Cold neutron sources at ILL. Nucl. Instrum. Methods A*, **284**, 197–199.

Akcasu, A. Z., Lellouche, G. S. & Shotkin, L. M. (1971). *Mathematical Methods in Nuclear Reactor Dynamics.* New York: Academic Press.

Alberi, J., Fischer, J., Radeka, V., Rogers, L. C. & Schoenborn, B. P. (1975). *A two-dimensional position-sensitive detector for thermal neutrons. Nucl. Instrum. Methods*, **127**, 507–523.

Alsmiller, R. G. & Lillie, R. A. (1992). *Design calculations for the ANS cold source. Part II. Heating rates. Nucl. Instrum. Methods A*, **321**, 265–270.

Bacon, G. E. (1962). *Neutron Diffraction.* Oxford University Press.

Böni, P. (1997). *Supermirror-based beam devices. Physica B*, **234–236**, 1038–1043.

Borkowski, C. J. & Kopp, M. K. (1975). *Design and properties of position-sensitive proportional counters using resistance-capacitance position encoding. Rev. Sci. Instrum.* **46**, 951–962.

Carpenter, J. M. (1977). *Pulsed spallation neutron sources for slow neutron scattering. Nucl. Instrum. Methods*, **145**, 91–113.

Carpenter, J. M. & Yelon, W. B. (1986). *Neutron sources.* In *Methods of Experimental Physics*, Vol. 23A. New York: Academic Press.

Cipriani, F., Castagna, J.-C., Caustre, L., Wilkinson, C. & Lehmann, M. S. (1997). *Large area neutron and X-ray image-plate detectors for macromolecular biology. Nucl. Instrum. Methods A*, **392**, 471–474.

Clark, C. D., Mitchell, E. W. J., Palmer, D. W. & Wilson, I. H. (1966). *The design of a velocity selector for long wavelength neutrons. J. Sci. Instrum.* **43**, 1–5.

Convert, P. & Forsyth, J. B. (1983). Editors. *Position-Sensitive Detection of Thermal Neutrons.* London: Academic Press.

Copley, J. R. D. (1991). *Acceptance diagram analysis of the performance of vertically curved neutron monochromators. Nucl. Instrum. Methods*, **301**, 191–201.

Copley, J. R. D. & Mildner, D. F. R. (1992). *Simulation and analysis of the transmission properties of curved–straight neutron guide systems. Nucl. Sci. Eng.* **110**, 1–9.

Crawford, R. K. (1992). *Position-sensitive detection of slow neutrons – survey of fundamental principles. SPIE*, **1737**, 210–223.

Ebisawa, T., Achiwa, N., Yamada, S., Akiyoshi, T. & Okamoto, S. (1979). *Neutron reflectivities of Ni–Mn and Ni–Ti multilayers for monochromators and supermirrors. J. Nucl. Sci. Technol.* **16**, 647–659.

Freund, A. K. & Dolling, G. (1995). *Devices for neutron beam definition.* In *International Tables for Crystallography*, Vol. C. *Mathematical, Physical and Chemical Tables*, edited by A. J. C. Wilson, pp. 375–382. Dordrecht: Kluwer Academic Publishers.

Glasstone, S. & Sesonske, A. (1994). *Nuclear Reactor Engineering.* New York: Chapman and Hall.

Hallsall, M. J. (1995). *WIMS – a general purpose code for reactor core analysis.* AEA Technology, Vienna.

Harris, P., Lebech, B. & Pedersen, J. S. (1995). *The three-dimensional resolution function for small-angle scattering and Laue geometries. J. Appl. Cryst.* **28**, 209–222.

Hayter, J. B. & Mook, H. A. (1989). *Discrete thin-film multilayer design for X-ray and neutron supermirrors. J. Appl. Cryst.* **22**, 35–41.

Hjelm, R. (1996). Editor. *Proceedings of the Workshop on Methods for Neutron Scattering Instrumentation Design.* Lawrence Berkeley National Laboratory, USA.

Hughes, H. G. III (1988). *Monte Carlo simulation of the LANSCE target geometry*. *Proceedings of the Tenth International Collaboration on Advanced Neutron Sources*, p. 455. New York: Institute of Physics.

Jacobé, J., Feltin, D., Rambaud, A., Ratel, F., Gamon, M. & Pernock, J. B. (1983). *High pressure ^{3}He multielectrode detectors for neutron localisation.* In *Position-Sensitive Detection of Thermal Neutrons*, edited by P. Convert & J. B. Forsyth, pp. 106–119. London: Academic Press.

Jakeman, D. (1966). *Physics of Nuclear Reactors.* London: The English Universities Press.

Johnson, M. W. (1986). Editor. *Workshop on Neutron Scattering Data Analysis.* Rutherford Appleton Laboratory, Chilton, England. Bristol: Institute of Physics.

Johnson, M. W. & Stephanou, C. (1978). *MCLIB: a library of Monte Carlo subroutines for neutron scattering problems.* Report RL-78–090. Science Research Council, Chilton, England.

Knott, R. B., Smith, G. C., Watt, G. & Boldeman, J. B. (1997). *A large 2D PSD for thermal neutron detection. Nucl. Instrum. Methods A*, **392**, 62–67.

Komura, S., Takeda, T., Fujii, H., Toyoshima, Y., Osamura, K., Mochiki, K. & Hasegawa, K. (1983). *The 6-meter neutron small-angle scattering spectrometer at KUR. Jpn. J. Appl. Phys.* **22**, 351–356.

Kostorz, G. (1979). *Neutron Scattering. Treatise on Materials Science and Technology*, Vol. 15. New York: Academic Press.

Krueger, S., Koenig, B. W., Orts, W. J., Berk, N. F., Majkrzak, C. F. & Gawrisch, K. (1996). *Neutron reflectivity studies of single lipid bilayers supported on planar substrates.* In *Neutrons in Biology*, edited by B. P. Schoenborn & R. B. Knott, pp. 205–213. New York: Plenum Press.

Lewis, E. E. & Miller, W. F. (1993). *Computational Methods of Neutron Transport.* Washington: American Nuclear Society Inc.

Lillie, R. A. & Alsmiller, R. G. (1990). *Design calculations for the ANS cold neutron source. Nucl. Instrum. Methods A*, **295**, 147–154.

Lowde, R. D. (1960). *The principles of mechanical neutron-velocity selection. J. Nucl. Energy*, **11**, 69–80.

Mâaza, M., Farnoux, B., Samuel, F., Sella, C., Wehling, F., Bridou, F., Groos, M., Pardo, B. & Foulet, G. (1993). *Reduction of the interfacial diffusion in Ni–Ti neutron-optics multilayers by carburation of the Ni–Ti interfaces. J. Appl. Cryst.* **26**, 574–582.

Magerl, A. & Wagner, V. (1994). Editors. *Proceedings of the workshop on focusing Bragg optics. Nucl. Instrum. Methods A*, Vol. 338.

Maier-Leibnitz, H. & Springer, T. (1963). *The use of neutron optical devices on beam-hole experiments. J. Nucl. Energy*, **17**, 217–225.

Majkrzak, C. F. (1991). *Polarised neutron reflectometry. Physica B*, **173**, 75–88.

Mikula, P., Krüger, E., Scherm, R. & Wagner, V. (1990). *An elastically bent silicon crystal as a monochromator for thermal neutrons. J. Appl. Cryst.* **23**, 105–110.

Mildner, D. F. R. & Hammouda, B. (1992). *The transmission of curved neutron guides with non-perfect reflectivity. J. Appl. Cryst.* **25**, 39–45.

Niimura, N., Karasawa, Y., Tanaka, I., Miyahara, J., Takahashi, K., Saito, H., Koizumi, S. & Hidaka, M. (1994). *An imaging plate neutron detector. Nucl. Instrum. Methods A*, **349**, 521–525.

Niimura, N., Minezaki, Y., Nonaka, T., Castagna, J.-C., Cipriani, F., Høghøj, P., Lehmann, M. S. & Wilkinson, C. (1997). *Neutron Laue diffractometry with an imaging plate provides an effective data collection regime for neutron protein crystallography. Nature Struct. Biol.* **4**, 909–914.

Oed, A. (1988). *Position-sensitive detector with microstrip anode for electron multiplication with gases. Nucl. Instrum. Methods A*, **263**, 351–359.

Oed, A. (1995). *Properties of micro-strip gas chambers (MSGC) and recent developments. Nucl. Instrum. Methods A*, **367**, 34–40.

Pedersen, J. S., Posselt, D. & Mortensen, K. (1990). *Analytical treatment of the resolution function for small-angle scattering. J. Appl. Cryst.* **23**, 321–333.

Popovici, M. & Yelon, W. B. (1995). *Focusing monochromators for neutron diffraction. J. Neutron Res.* **3**, 1–26.

Prael, R. E. (1994). *A review of the physics models in the LAHET code*. Report LA-UR-94-1817. Los Alamos National Laboratory, USA.

Prask, H. J., Rowe, J. M., Rush, J. J. & Schroeder, I. G. (1993). *The NIST cold neutron research facility. J. Res. NIST*, **98**, 1–14.

Pynn, R. (1984). *Neutron scattering instrumentation at reactor based installations. Rev. Sci. Instrum.* **55**, 837–848.

Radeka, V. (1988). *Low noise techniques in detectors. Annu. Rev. Nucl. Part. Sci.* **38**, 217–277.

Radeka, V. & Boie, R. A. (1980). *Centroid finding method for position-sensitive detectors. Nucl. Instrum. Methods*, **178**, 543–554.

Radeka, V., Schaknowski, N. A., Smith, G. C. & Yu, B. (1996). *High precision thermal neutron detectors.* In *Neutrons in Biology*, edited by B. P. Schoenborn & R. B. Knott, pp. 57–67. New York: Plenum Press.

Rausch, C., Bücherl, T., Gähler, R., Seggern, H. & Winnacker, A. (1992). *Recent developments in neutron detection. SPIE*, **1737**, 255–263.

Richter, D. & Springer, T. (1998). *A Twenty Years Forward Look at Neutron Scattering Facilities in the OECD Countries and Russia*. OECD Publication. Strasbourg: European Science Foundation.

Riste, T. (1970). *Singly bent graphite monochromators for neutrons. Nucl. Instrum. Methods.* **86**, 1–4.

Russell, G. J., Ferguson, P. D., Pitcher, E. J. & Court, J. D. (1996). *Neutronics and the MLNSC spallation target system*. In *Applications of Accelerators in Research and Industry – Proceedings of the 14th International Conference*, edited by J. L. Duggan and I. L. Morgan. AIP Conference Proceedings, Vol. 392, pp. 361–364.

Sauli, F. (1977). *Principles of operation of multiwire proportional and drift chambers.* Report CERN-77-09. CERN, Geneva, Switzerland.

Saxena, A. M. & Schoenborn, B. P. (1977). *Multilayer neutron monochromators. Acta Cryst.* A**33**, 805–813.

Saxena, A. M. & Schoenborn, B. P. (1988). *Multilayer monochromators for neutron spectrometers. Mater. Sci. Forum*, **27/28**, 313–318.

Schärpf, O. & Anderson, I. S. (1994). *The role of surfaces and interfaces in the behaviour of non-polarizing and polarizing supermirrors. Physica B*, **198**, 203–212.

Schefer, J., Medarde, M., Fischer, S., Thut, R., Koch, M., Fischer, P., Staub, U., Horisberger, M., Bottger, G. & Donni, A. (1996). *Sputtering method for improving neutron composite germanium monochromators. Nucl. Instrum. Methods A*, **372**, 229–232.

Schneider, D. K. & Schoenborn, B. P. (1984). *A new neutron small-angle diffraction instrument at the Brookhaven High Flux Beam Reactor.* In *Neutrons in Biology*, edited by B. P. Schoenborn, pp. 119–141. New York: Plenum Press.

Schoenborn, B. P. (1992*a*). *Multilayer monochromators and super mirrors for neutron protein crystallography using a quasi Laue technique. SPIE*, **1738**, 192–199.

Schoenborn, B. P. (1992*b*). *Area detectors for neutron protein crystallography. SPIE*, **1737**, 235–243.

Schoenborn, B. P. (1996). *A protein crystallography station at the Los Alamos Neutron Science Center*. Report LA-UR-96-3508, 11–64. Los Alamos National Laboratory, USA.

Schoenborn, B. P., Court, D., Larson, A. C. & Ferguson, P. (1999). *Moderator decoupling options for structural biology at spallation neutron sources. J. Neutron Res.* **7**, 89–106.

Schoenborn, B. P., Saxena, A. M., Stamm, M., Dimmler, G. & Radeka, V. (1985). *A neutron spectrometer with a two-dimensional detector for time resolved studies. Aust. J. Phys.* **38**, 337–351.

Schoenborn, B. P., Schefer, J. & Schneider, D. (1986). *The use of wire chambers in structural biology. Nucl. Instrum. Methods A*, **252**, 180–187.

Sears, V. F. (1983). *Theory of multilayer neutron monochromators. Acta Cryst.* A**39**, 601–608.

Sears, V. F. (1989). *Neutron Optics: an Introduction to the Theory of Neutron Optical Phenomena and their Applications. Oxford Series on Neutron Scattering in Condensed Matter.* New York: Oxford University Press.

Sivia, D. S., Silver, R. N. & Pynn, R. (1990). *The Bayesian approach to optimal instrument design.* In *Neutron Scattering Data Analysis*, edited by M. W. Johnson, Institute of Physics Conference Series, Vol. 107, pp. 45–55.

Soodak, H. (1962). Editor. *Reactor Handbook.* New York: Wiley.

Spanier, J. & Gelbard, E. M. (1969). *Monte Carlo Principles and Neutron Transport Problems.* London: Addison-Wesley.

Stamm'ler, R. J. J. & Abbate, M. J. (1983). *Methods of Steady-State Reactor Physics in Nuclear Design.* London: Academic Press.

Stuhrmann, H. B. & Nierhaus, K. H. (1996). *The determination of the in situ structure by nuclear spin contrast variation*. In *Neutrons in Biology*, edited by B. P. Schoenborn & R. B. Knott, pp. 397–413. New York: Plenum Press.

Takahashi, K., Tazaki, S., Miyahara, J., Karasawa, Y. & Niimura, N. (1996). *Imaging performance of imaging plate neutron detectors. Nucl. Instrum. Methods A*, **377**, 119–122.

Vellettaz, N., Assaf, J. E. & Oed, A. (1997). *Two dimensional gaseous microstrip detector for thermal neutrons. Nucl. Instrum. Methods A*, **392**, 73–79.

Vogt, T., Passell, L., Cheung, S. & Axe, J. D. (1994). *Using wafer stacks as neutron monochromators. Nucl. Instrum. Methods A*, **338**, 71–77.

Wagner, V., Friedrich, H. & Wille, P. (1992). *Performance of a high-tech neutron velocity selector. Physica B*, **180–181**, 938–940.

Weisman, J. (1983). Editor. *Elements of Nuclear Reactor Design*. Amsterdam: Elsevier Scientific Publishing Company.

Well, A. A. van, de Haan, V. O. & Mildner, D. F. R. (1991). *The average number of reflections in a curved neutron guide. Nucl. Instrum. Methods A*, **309**, 284–286.

West, C. D. (1989). *The US Advanced Neutron Source*. ICANS X, Los Alamos USA, pp. 643–654.

Wignall, G. D., Christen, D. K. & Ramakrishnan, V. (1988). *Instrumental resolution effects in small-angle neutron scattering. J. Appl. Cryst.* **21**, 438–451.

Williams, M. M. R. (1966). *The Slowing Down and Thermalization of Neutrons*. Amsterdam: North Holland.

Windsor, C. G. (1981). *Pulsed Neutron Scattering*. London: Wiley.

Windsor, C. G. (1986). *Experimental techniques*. In *Methods of Experimental Physics*, Vol. 23A. New York, London: Academic Press.

International Tables for Crystallography (2012). Vol. F, Chapter 7.1, pp. 177–182.

PART 7. X-RAY DETECTORS

Chapter 7.1. Comparison of X-ray detectors

S. M. Gruner, E. F. Eikenberry and M. W. Tate

7.1.1. Commonly used detectors: general considerations

This chapter summarizes detector characteristics and provides practical advice on the selection of crystallographic detectors. Important types of detectors for crystallographic applications are summarized in Section 7.1.2 and listed in Table 7.1.1.1. To be detected by any device, an X-ray must be absorbed within a detective medium through electrodynamic interactions with the atoms in the detecting layer. These interactions usually result in an energetic electron being liberated which, by secondary and tertiary interactions, produces the signal that will be measured, *e.g.* luminescence in phosphors, electron–hole pairs in semiconductors or ionized atoms in gaseous ionization detectors. As we shall see, there are many schemes for recording these signals. The various detector designs, as well as the fundamental detection processes, have particular advantages and weaknesses. In practice, detector suitability is constrained by the experimental situation (*e.g.* home laboratory X-ray generator *versus* synchrotron-radiation source; fine-slicing *versus* large-angle oscillations), by the sample (*e.g.* whether radiation damages it readily) and by availability. An assessment of detector suitability in a given situation requires an understanding of how detectors are evaluated and characterized. Some of the more important criteria are discussed below.

The *detective quantum efficiency* (DQE) is an overall measure of the efficiency and noise performance of a detector (Gruner *et al.*, 1978). The DQE is defined as

$$\mathrm{DQE} = (S_o/N_o)^2/(S_i/N_i)^2, \qquad (7.1.1.1)$$

where S is the signal, N is the noise, and the subscripts o and i refer to the output and input of the detector, respectively. The DQE measures the degradation owing to detection in the signal-to-noise ratio. For a signal source that obeys Poisson statistics, the inherent noise is equal to the square root of the number of incident photons, so that the incident signal-to-noise ratio is just $S_i/N_i = (S_i)^{1/2}$. The ideal detector introduces no additional noise in the detection process, thereby preserving the incident signal-to-noise ratio, *i.e.* $\mathrm{DQE} = 1$. Real detectors always have $\mathrm{DQE} < 1$ because some noise is always added in the detection process. The DQE automatically accounts for the fact that the input and output signals may be of a different nature (*e.g.* X-rays in, stored electrons out), since it is a ratio of dimensionless numbers.

A single number does not characterize the DQE of a system. Rather, the DQE is a function of the integrated dose, the X-ray spot size, the length of exposure, the rate of signal accumulation, the X-ray energy *etc.* Noise in the detector system will limit the DQE at low dose, while the inability to remove all systematic nonuniformities will limit the high-dose behaviour.

The *accuracy*, ρ, measures the output noise relative to the signal, *i.e.* $\rho = N_o/S_o$. For a Poisson X-ray source, it follows that the accuracy and the DQE are related by

$$\rho = (N_i \mathrm{DQE})^{-1/2}. \qquad (7.1.1.2)$$

This allows the determination of the number of X-rays needed to measure a signal to a given accuracy with a detector of a given DQE. The accuracy for an ideal detector is $1/(S_i)^{1/2}$, *e.g.* 100 X-rays are required to measure to 10% accuracy, and 10^4 X-rays are needed for a 1% accuracy. Nonideal detectors ($\mathrm{DQE} < 1$) always require more X-rays than the ideal to measure to a given accuracy.

Spatial resolution refers to the ability of a detector to measure adjacent signals independently. The spatial resolution is characterized by the *point spread function* (PSF), which, for most detectors, is simply the spread of intensity in the output image as a result of an incident point signal. An alternative measure of resolution is the *line spread function* (Fujita *et al.*, 1992).

Table 7.1.1.1
X-ray detectors for crystallography

(*a*) Commercially available detectors

Technology	Primary X-ray converter	Format
Film	AgBr	Area
Storage phosphor	BaFBr	Area
Scintillating crystal	NaI, CsI	Point
Gas discharge	Xe	Point, linear, area
Television	Phosphor	Area
CCD	Phosphor	Area
Silicon diode	Si	Linear, area
Avalanche diode	Si	Point, area

(*b*) Detectors under development

Technology	Primary X-ray converter	Format
Pixel array	Si, GaAs, CdZnTe	Area
Amorphous silicon flat panel + phosphor	CsI, Gd_2O_2S	Area
Amorphous silicon flat panel + photoconductor	PbI_2, CdZnTe, TlBr, HgI_2	Area

Copyright © 2012 International Union of Crystallography

Although a detector might have a narrow PSF at 50% of the peak level, poor performance of the PSF at the 1% level and below can severely hamper the ability to measure closely spaced spots. It is important to realize that the PSF is a two-dimensional function, which is often illustrated by a graph of the PSF cross section; therefore, the integrated intensity at a radius R pixels from the centre of the PSF is the value of the PSF cross section times the number of pixels at that radius. Often the wings of the PSF decay slowly, so that considerable integrated signal is in the image far from the spot centre. In this case, a bright spot can easily overwhelm a nearby weak spot. Another consequence is that bright spots appear considerably larger than dim ones, thereby complicating analysis.

The *stopping power* is the fraction of the incident X-rays that are stopped in the active detector recording medium. In low-noise detectors, the DQE is proportional to the stopping power. A detector with low stopping power may be suitable for experiments in which there is a strong X-ray signal from a specimen that is not readily damaged by radiation. On the other hand, even a noiseless detector with a low stopping power will have a low DQE, because most of the incident X-rays are not recorded.

Unfortunately, many definitions of *dynamic range* are used for detectors. For an integrating detector, the dynamic range per pixel is taken to be the ratio of the saturation signal per pixel to the zero-dose noise per pixel for a single frame readout. For photon counters, the dynamic range per pixel refers to the largest signal-to-noise ratio, *i.e.* the number of true counts per pixel that are accumulated on average before a false count is registered. In practice, the dynamic range is frequently limited by the readout apparatus or the reproducibility of the detector medium. For example, the large dynamic range of storage phosphors is almost always limited by the capabilities of the reading apparatus, which constrains the saturation signal and limits the zero-dose noise by the inability to erase the phosphor completely. The number of bits in the output word does not indicate the dynamic range, since the number of stored bits can only constrain the dynamic range, but, obviously, cannot increase it.

The dynamic range is sometimes given with respect to an integrated signal that spans more than one pixel. For a signal S per pixel which spans M pixels, the integrated signal is MS, and, assuming the noise adds in quadrature, the noise is $N(M)^{1/2}$, yielding a factor of $(M)^{1/2}$ larger dynamic range. For most detectors, the noise in nearby pixels does not add in quadrature, so this is an upper limit.

The characteristics of a detector may be severely compromised by practical considerations of *nonlinearity*, *reproducibility* and *calibration*. For example, the optical density of X-ray film varies nonlinearly with the incident dose. Although it is possible to calibrate the optical density *versus* dose response, in practice it is difficult to reproduce exactly the film-developing conditions required to utilize the highly nonlinear portions of the response. A detector is no better than its practical calibration. This is especially true for area detectors in which the sensitivity varies across the face of the detector. The proper calibration of an area detector is replete with subtleties and constrained by the long-term stability of the calibration. Faulty calibrations are responsible for much of the difference between the possible and actual performance of detectors (Barna *et al.*, 1999).

The response of a detector may be nonlinear with respect to position, dose, intensity and X-ray energy. Nonuniformity of response across the active area is compensated by the *flat-field* correction. Frequently, nonuniformity of response varies with the angle of incidence of the X-ray beam to the detector surface, which is a significant consideration when flat detectors are used to collect wide-angle data. Although this may be compensated by an energy-dependent *obliquity* correction, few detector vendors provide this calibration. An X-ray image may also be spatially distorted; this *geometric distortion* can be calibrated if it is stable.

Other important detector considerations include the *format* of the detector (*e.g.* the number of pixels across the height and width of the detector). The format and the PSF together determine the number of Bragg orders that can be resolved across the active area of the detector. *Robustness* of the detector is also important: as examples, gas-filled area detectors may be sensitive to vibration of the high-voltage wires; detectors containing image intensifiers are sensitive to magnetic fields; or the detector may simply be easily damaged or lose its calibration during routine handling. Some detectors are readily damaged by too large an X-ray signal. *Count-rate* considerations severely limit the use of many photon counters, especially at synchrotron-radiation sources. Detector *speed*, both during exposure and during read out, can be important. Some detector designs are highly *flexible*, permitting special readout modes, such as a selected region of interest for use during alignment, or operation as a streak camera.

Ease of use is especially important. A detector may simply be hard to use because, for example, it is exceptionally delicate, requires frequent fills of liquid nitrogen, or is physically awkward in size. A final, often compelling, consideration is whether a detector is *well integrated into an application* with the appropriate analysis software and whether the control software is well interfaced to the other X-ray hardware.

7.1.2. Evaluating and comparing detectors

The DQE comprehensively characterizes the ultimate quantitative capabilities of an X-ray detector. The DQE may be determined from an analysis of the reproducibility of recorded X-ray test images of known statistics *via* equation (7.1.1.1): given M incident X-rays per exposure, the expected incident signal-to-noise is $(M)^{1/2}$. The DQE is determined by measuring the variance in the recorded signal in repeated measurements of the test image. Repetition of this process for different values of M maps out the DQE curve. Since the DQE is dependent on the structure of the image, the integration area, the X-ray background and the long-term detector calibration, it is essential that the test images realistically simulate these features as expected in experiments. Thus, if the detector is to be used to obtain images of diffraction spots, the test images should consist of comparably sized spots superimposed on a suitable background.

A comprehensive DQE determination is nontrivial and requires specialized tools, such as test masks, uniform X-ray sources *etc.* Unfortunately, published DQE curves are frequently incorrect and misleading. Users can, however, set up and perform a simple DQE assessment, detailed below, which gives a great deal of information about the sensitivity and usefulness of a given area detector. Other sources of stable X-ray spots (of appropriate size and intensity) can also be used in similar tests.

The materials needed are sheet lead and aluminium, a sewing needle, a stable collimated X-ray source, X-ray capillaries filled with saturated salt solutions, an X-ray shutter with timing capability and a scintillator/phototube X-ray counting arrangement. Arrange a fluorescent X-ray source to provide a diffuse X-ray signal. An X-ray capillary filled with a saturated solution of iron chloride makes a suitable source for a copper anode machine. Next, make an X-ray-opaque metal mask by punching a clean pinhole with a sewing needle in a lead sheet. The size of the hole should be representative of an X-ray spot, say 0.3 mm in diameter. The mask should be firmly and reproducibly secured a few cm from the fluorescent source at a wide angle to the incident beam. Using a scintillator/phototube combination, measure the number of X-rays per second emerging through the hole at a given X-ray source loading. A sufficient number of X-rays per measurement (say 10^5) is necessary to obtain accurate statistics (0.3%). This measurement should be repeated to verify the stability of the source.

This spot can now be recorded by the detector in question, using different integration times to vary the dose. 20 measurements at each integration time should give a reliable measure of the standard deviation in the signal. It is vital to move the position of the spot on the detector face for each exposure, taking care to move only the detector without disturbing the remainder of the experimental setup. Only by moving the detector is the fidelity of the calibrations tested. One subtlety is that the sensitivity of many detectors varies with the angle of incidence of the X-rays, so that it will be necessary to vary both the position and angle of the detector between exposures.

By using a wide range of integration times, both the sensitivity of the detector at low doses and the ultimately achievable measurement accuracy can be examined. These data may also highlight specific problems a detector might have, such as nonlinearity.

The DQE can be measured for a spot in the presence of a background if the lead pinhole mask is now replaced with a pinhole in a semitransparent aluminium foil. Choose the foil thickness to yield an appropriate background level, say 20% of the pinhole intensity. The uncertainty in the measurement of the spot intensity now results from the total counts in the integration area in addition to the uncertainty in determining the background. A wide PSF is especially harmful in this case, since many more pixels must be integrated to encompass the spot.

These evaluation procedures test only limited aspects of the detector, but in doing so, much is learned not only about the detector, but also about the degree to which the vendor is willing to work with the user, which is clearly of interest. The ultimate test for a crystallographer is whether a detector delivers good data in a well understood experimental protocol. Usually, values of $R_{\rm sym}$, the agreement of integrated intensities from symmetry-related reflections, are evaluated as a function of resolution. Low values of $R_{\rm sym}$ suggest good quality data. A much more stringent test can be made by comparing anomalous difference Patterson maps based on the Fe atom in myoglobin (Krause & Phillips, 1992). The limitation in these crystallography-based evaluations is that they tend to rely on robust, strongly diffracting crystals, which allow accumulation of good X-ray statistics even with insensitive detectors. Weakly diffracting and radiation-sensitive crystals are less forgiving.

7.1.3. Characteristics of different detector approaches

7.1.3.1. Point versus linear versus area detection

A point detector may be based on a scintillating crystal or a gas-filled counter, with the sensitive area defined by slits or a pinhole mask. The spatial resolution of such a detector can be made arbitrarily fine at the expense of data collection rate. Point detectors can have very high accuracy if the background is removed by energy discrimination. They find application in powder diffractometry and small-molecule crystallography, in which the reflections are widely dispersed, thereby simplifying measurement of individual reflections. Clearly, specimen and source stability are important for such work.

Throughput can be greatly increased by area detection, which is often required for macromolecular crystallography or investigations of unstable specimens. Typical area detectors, such as film, storage phosphors and charge-coupled devices (CCDs), are described below.

7.1.3.2. Counting and integrating detectors

Detectors can be broadly divided into photon counters and photon integrators. Photon counters have the advantage that some designs permit energy discrimination, allowing them to reject inelastically scattered radiation, thereby improving the signal-to-noise ratio. However, photon-counting detectors always have a count-rate limitation, above which they begin to miss events, or even become unresponsive (the time during which a detector misses events is known as dead time). Prototype systems have demonstrated linear count rates greater than 10^6 photon s^{-1}. Fabrication difficulties have limited the commercial availability of photon-counting detectors with large areas, high spatial resolution and high count rates. The count rate is a particular concern at modern synchrotron sources, which are capable of generating diffraction that delivers two or more photons to a pixel during one bunch time, an instantaneous count rate greater that 10^{10} photons per second per pixel. Integrating detectors are more typically used in situations where very high event rates are expected.

In contrast, integrating detectors have no inherent count-rate limitation, though at very high fluxes several sources of nonlinearity can theoretically become important, such as nonlinearity in the phosphor used to convert the X-ray image to a visible image. Integrating detectors, however, do not discriminate energy, and they have noise that increases with integration time. Nonetheless, film, image-plate and CCD integrating detectors are currently commercially available and in widespread use.

7.1.3.2.1. Photon-counting detectors

Commonly used photon counters include *scintillator/photomultiplier combinations*, *gas-filled counters* and *reverse-biased semi-conductor detectors*.

Scintillator/photomultipliers usually consist of a relatively thick crystal of a scintillator coupled to a high-gain photomultiplier tube. These detectors are generally designed to serve as point photon counters with moderate energy resolution. In order to perform this function, several constraints must be met:

(1) The scintillator crystal must be thick enough to have almost unity stopping power.

(2) It is necessary to collect as many of the converted visible photons as possible, so an optically clean scintillator crystal is used in a reflective housing to direct as many photons as possible toward the phototube.

(3) The scintillator must emit its light quickly, so as to minimize dead time, and be efficient, so as to emit much light. NaI:Tl, CsI:Na and CsI:Tl meet these constraints. NaI is more commonly used, but CsI may be preferred at higher X-ray energies because of its higher stopping power. Both materials are hygroscopic and are usually encased in hermetically sealed capsules with beryllium windows.

(4) The phototube is usually operated in its linear region for energy discrimination.

Scintillator/phototube combinations are relatively trouble-free and often have near-unity DQE. Their main limitations are count rates well below 10^6 photon s^{-1} and the lack of spatial resolution. Even so, such detectors are still preferred in many applications where the data are effectively zero- or one-dimensional.

Reverse-biased semiconductor detectors are designed to have a thick depletion zone in which charge can be efficiently collected and conveyed to an amplifier. X-rays that stop in the depletion zone produce electron–hole pairs; these are separated by the depletion zone field and the electrons are swept to the input of a low-noise amplifier. Single-photon counting can be readily achieved, even for low-energy X-rays, especially if the detector is cooled to minimize thermally generated charge. These detectors are typically fabricated as silicon diodes, but germanium and gallium arsenide are also used (Hall, 1995). Until recently, these devices were generally configured as point detectors or strip detectors consisting of a linear array of narrow sensitive regions, forming a one-dimensional detector (Ludewigt *et al.*, 1994). Two-dimensional arrays of square pixels are being developed, *e.g.* see the description of pixel array detectors below. In another device, the silicon drift detector, potentials are arranged in the silicon to funnel signals from a large area to a low-noise collection point (Rehak *et al.*, 1986). Such devices are being developed for both linear and area applications.

By increasing the electric field strength in an appropriately designed p-n junction in silicon, avalanche multiplication of the X-ray-induced electrons can be obtained as they move toward the anode where they are collected. This gives rise to a very high linear signal gain, with high speed and low noise. Arrays of such *avalanche photodiodes* as large as 8 × 8 elements, each 1 × 1 mm square, have been fabricated (Gramsch *et al.*, 1994; Farrell *et al.*, 1994).

Gas discharge (wire) counters make use of the ionization produced when an X-ray is stopped in the high-atomic-number gas, usually xenon, that fills the detector. A strong electric field between a fine anode wire and a cathode plane accelerates the products of the primary ionization to produce an ionizing multiplication (either a proportional or an avalanche discharge, depending on the field strength) that is detected as a charge pulse on one or both of the electrodes. The discharge is quenched by the presence of a few per cent of a second gas, *e.g.* methane or carbon dioxide. Gas discharge detectors have been configured in zero-, one- and two-dimensional versions and continue to be widely used in some applications. The venerable Geiger counter is in this class and is used for radiation monitoring and beam alignment in home laboratories. Properly designed gas discharge counters have very low noise, but the quantum efficiency depends critically on design, gas and X-ray energy.

Linear wire detectors have been used to record small-angle X-ray scattering. The localization of the X-ray event along the length of the detector is often performed by measuring the difference in arrival time of the charge pulses at the two ends of one of the electrodes (Barbosa *et al.*, 1989). The pulses are stretched to permit this measurement. One design uses a resistive anode wire to perform this function, whereas others configure the cathode plane as a delay line. Various two-dimensional arrangements of crossed planes of wires, broadly classified as multiwire proportional counters (MWPCs), have been widely used in crystallography, and some types have been commercially successful (Hamlin *et al.*, 1981; Blum *et al.*, 1987).

The design of MWPC area detectors has had difficulty keeping up with improvements in X-ray sources, particularly the high fluxes available at storage rings, and the shift toward use of higher-energy X-rays. The electric discharge at the heart of the technology has an inherent dead time associated with it. Added to this inherent dead time are the pulse propagation and processing times which limit the counting rate for a given wire. Thus, MWPCs are subject to a severe count-rate limitation. A second limitation of MWPCs has been their large pixel size and the relatively small number of pixels across the detector face, as well as parallax effects. These problems have been addressed by changes in the detector geometry (*e.g.* spherical drift chambers; Charpak, 1982), by microfabrication on glass substrates of the wires comprising the back plane of the detector, and by dividing the active area into small zones, each of which is read out independently. Robustness of MWPCs has also been a problem.

The dead time can be reduced by reducing the thickness of the detector. However, reducing the detector thickness reduces the X-ray stopping power. Increasing the gas pressure not only improves the quantum efficiency, but also helps to reduce the dead time further. Unfortunately, high gas pressure complicates the design of the front window of the detector. Despite these problems, two-dimensional gas-detector prototype modules with 200 μm square pixels have been constructed that are expected to have a local linear count-rate limit of 7 MHz mm^{-2} and a quantum efficiency above 80% at energies used in crystallography (see Sarvestani *et al.*, 1998).

7.1.3.2.2. Integrating detectors

X-ray film was the first area detector and has a long history of important contributions to the solution of structures and to X-ray imaging (Arndt *et al.*, 1977). In many applications, film has effectively been displaced because of its relative insensitivity caused by the high level of background fog, its multistep processing leading to long delays before digital data are obtained and its nonlinearity. However, X-ray film is still a superior integrator for long exposures, and lithographic film has much higher spatial resolution than any other area detector. Polaroid film in an X-ray cassette is an excellent diagnostic tool for beam alignment problems. And, in all cases, film is inexpensive.

Storage phosphors, also called image plates, are probably the most widely used area X-ray detector for crystallography at present, particularly in laboratories with conventional X-ray sources (Amemiya *et al.*, 1988; Eikenberry *et al.*, 1992). These sheets of material are a much improved functional replacement for X-ray film in many applications, including medical radiography and autoradiography in biological research. Storage phosphors are made from a BaFBr:Eu or other photostimulable phosphor coated on a suitable backing. These phosphors have the property that absorbed X-ray energy can be trapped in long-lived states within the phosphor grains, and that this energy can later be released as blue fluorescence upon photostimulation with red light. Grain-size distribution, grain orientation, binder choice and coating thickness are important parameters in the commercial preparation of the sheets.

The exposed phosphor sheet is raster-scanned with a finely focused red laser and the resulting photostimulated emission is recorded by a photomultiplier. The result is a digital image of the X-ray intensity distribution. The scanning can be done either on-line in self-contained systems or off-line in a separate scanning instrument. Scanning typically requires several minutes. In advanced scanners, this is reduced to several tens of seconds, where the limit is set by the time constant of the photostimulated emission process, which in turn determines the minimum time the laser should dwell on each pixel. The off-line scanner, preferred at synchrotron sources, permits a new exposure to be made while scanning is performed. Self-contained systems offer the advantages of simplified operation and the possibility of calibrating the detector, since only a single sheet of material is used in a mechanically stable setup.

Storage phosphors are typically read out with 100 µm square pixels, resulting in 2000 × 2500 pixel images for the common size of sheet. Larger formats are available. The wide PSF of storage phosphors makes the effective pixel size considerably larger than the nominal value. Some scanners permit smaller pixels, but this is of limited utility because there is a readout noise component associated with each pixel and too small a pixel harms the signal-to-noise ratio without improving data. The most critical component of the storage-phosphor system is the mirror assembly that gathers the photostimulated emission during readout. There are very few emitted photons for each stored X-ray (at the photon energies used for diffraction), and only a small fraction of these are detected in the photomultiplier (in some cases less than one per stored X-ray). This step in the detection process is critical in maintaining high quantum efficiency. Although image plates have an inherently wide dynamic range, the practical value is always limited by the scanner analogue-to-digital converter.

Television detectors. Numerous integrating detector designs based on television sensor technologies have been published and, in one case, produced commercially (Milch *et al.*, 1982; Arndt, 1991). These detectors span a wide range of design complexity and performance. The primary element is a phosphor screen, which converts the incident X-ray pattern to a light image that is directly or indirectly coupled to the sensor, such as a Vidicon or CCD. Many of the designs employ image intensifiers to raise the signal strength of the visible image above the noise of the sensor system. Some designs employ cameras operated at video rates, with frames accumulated in the attached computer or on videotape. Other designs use cooled cameras operated in a slow-scan mode, which greatly reduces noise. The X-ray exposure is integrated in the camera, then read out once at the end of the integration period.

Most of these systems would be classified as complex, but several of them are working reliably today in laboratories with conventional X-ray sources. The image intensifiers improve the DQE of the systems, while sacrificing dynamic range and image sharpness. In addition, intensifiers are sensitive to magnetic fields, requiring great care in their use if proper detector calibration is to be maintained.

Considerable enhancement to the television-type detector is made possible by the low-noise imaging capabilities of CCDs, described in Chapter 7.2. In this case, high DQE can be maintained without intensification when the CCD is cooled and read with slow-scan electronics. As such, these detectors are much more robust and have improved imaging qualities.

7.1.4. Future detectors

Commercially available X-ray detectors have evolved from X-ray film and point diffractometry to area gas-proportional counters, to image plates, and now to CCD detectors. Two new X-ray detector technologies are on the horizon. One is based on the large-area amorphous semiconductors and thin-film transistor arrays which are being intensively developed by many large companies for medical radiography (reviewed by Moy, 1999). The radiographic need is to be able to cover very large areas (*e.g.* 0.5 m^2) with a high-spatial-resolution detector that is sensitive to hard X-rays. A number of these detectors are at the moment (1999) poised for introduction, but they are specialized for radiographic applications and are poorly suited for relatively long, low-noise integration of low-energy X-rays. It remains to be seen whether the technology will succeed and whether it can be modified for quantitative crystallographic applications.

A second technology being developed specifically for quantitative X-ray diffraction is based on solid-state pixel array detectors (PADs) (Iles *et al.*, 1996; Datte *et al.*, 1999; Barna *et al.*, 1997; Rossi *et al.*, 1999). In a PAD, X-rays are stopped directly in a semiconductor and the resulting signal is processed by electronics integrated into each pixel. Direct conversion of X-rays into electrical signals in a high-grade semiconductor has many advantages: many signal electrons are produced for each X-ray, and the conversion medium is very linear, has low noise and is well understood. Since each pixel has its own electronics, there is enormous flexibility in performing local signal processing. In principle, PADs have tremendous advantages of sensitivity, flexibility, noise and stability. The challenge will be to make PADs of a size and format useful for crystallography, while still being sufficiently affordable to be commercially viable.

References

Amemiya, Y., Matsushita, T., Nakagawa, A., Satow, Y., Miyahara, J. & Chikawa, J.-I. (1988). *Design and performance of an imaging plate system for X-ray diffraction study. Nucl. Instrum. Methods Phys. Res. A*, **266**, 645–653.

Arndt, U. W. (1991). *Second-generation X-ray television area detectors. Nucl. Instrum. Methods Phys. Res. A*, **310**, 395–397.

Arndt, U. W., Gilmore, D. J. & Wonacott, A. J. (1977). *X-ray film.* In *The Rotation Method in Crystallography*, edited by U. W. Arndt & A. J. Wonacott, pp. 207–218. Amsterdam: North-Holland Publishing Co.

Barbosa, A. F., Gabriel, A. & Craievich, A. (1989). *An X-ray gas position-sensitive detector – construction and characterization. Rev. Sci. Instrum.* **60**, 2315–2317.

Barna, S. L., Shepherd, J. A., Tate, M. W., Wixted, R. L., Eikenberry, E. F. & Gruner, S. M. (1997). *Characterization of prototype pixel array detector (PAD) for use in microsecond framing time-resolved X-ray diffraction studies. IEEE Trans. Nucl. Sci.* **44**, 950–956.

Barna, S. L., Tate, M. W., Gruner, S. M. & Eikenberry, E. F. (1999). *Calibration procedures for charge-coupled device X-ray detectors. Rev. Sci. Instrum.* **70**, 2927–2934.

Blum, M., Metcalf, P., Harrison, S. C. & Wiley, D. C. (1987). *A system for collection and on-line integration of X-ray diffraction data from a multiwire area detector. J. Appl. Cryst.* **20**, 235–242.

Charpak, G. (1982). *Parallax-free, high-accuracy gaseous detectors for X-ray and VUV localization. Nucl. Instrum. Methods*, **201**, 181–192.

Datte, P., Beuville, E., Millaud, J. & Xuong, N.-H. (1999). *A digital pixel address generator for pixel array detectors. Nucl. Instrum. Methods Phys. Res. A*, **421**, 492–501.

Eikenberry, E. F., Tate, M. W., Bilderback, D. H. & Gruner, S. M. (1992). *X-ray detectors: comparison of film, storage phosphors and CCD detectors. Inst. Phys. Conf. Ser.* **121**, 273–280.

Farrell, R., Vanderpuye, K., Cirignano, L., Squillante, M. R. & Entine, G. (1994). *Radiation detection performance of very high-gain avalanche photodiodes. Nucl. Instrum. Methods Phys. Res. A*, **353**, 176–179.

Fujita, H., Tsai, D.-Y., Itoh, T., Doi, K., Morishita, J., Ueda, K. & Ohtsuka, A. (1992). *A simple method for determining the modulation transfer-function in digital radiography. IEEE Trans. Med. Imaging*, **11**, 34–39.

Gramsch, E., Szawlowski, M., Zhang, S. & Madden, M. (1994). *Fast, high-density avalanche photodiode-array. IEEE Trans. Nucl. Sci.* **41**, 762–766.

Gruner, S. M., Milch, J. R. & Reynolds, G. T. (1978). *Evaluation of area photon detectors by a method based on detective quantum efficiency (DQE). IEEE Trans. Nucl. Sci.* NS-**25**, 562–565.

Hall, G. (1995). *Silicon pixel detectors for X-ray diffraction studies at synchrotron sources. Q. Rev. Biophys.* **28**, 1–32.

Hamlin, R., Cork, C., Howard, A., Nielsen, C., Vernon, W., Matthews, D. & Xuong, N. H. (1981). *Characteristics of a flat multiwire area detector for protein crystallography. J. Appl. Cryst.* **14**, 85–93.

Iles, G., Raymond, M., Hall, G., Lovell, M., Seller, P. & Sharp, P. (1996). *Hybrid pixel detector for time resolved X-ray diffraction experiments at synchrotron sources*. *Nucl. Instrum. Methods Phys. Res. A*, **381**, 103–111.

Krause, K. L. & Phillips, G. N. Jr (1992). *Experience with commercial area detectors: a 'buyer's' perspective. J. Appl. Cryst.* **25**, 146–154.

Ludewigt, B., Jaklevic, J., Kipnis, I., Rossington, C. & Spieler, H. (1994). *A high-rate, low-noise, X-ray silicon strip detector system*. *IEEE Trans. Nucl. Sci.* **41**, 1037–1041.

Milch, J. R., Gruner, S. M. & Reynolds, G. T. (1982). *Area detectors capable of recording X-ray diffraction patterns at high count rates*. *Nucl. Instrum. Methods*, **201**, 43–52.

Moy, J.-P. (1999). *Large area X-ray detectors based on amorphous silicon detector*. *Thin Solid Films*, **337**, 213.

Rehak, P., Walton, J., Gatti, E., Longoni, A., Sanpietro, M., Kemmer, J., Dietl, H., Holl, P., Klanner, R., Lutz, G., Wylie, A. & Becker, H. (1986). *Progress in semiconductor drift detectors*. *Nucl. Instrum. Methods Phys. Res. B*, **248**, 367–378.

Rossi, G., Renzi, M., Eikenberry, E. F., Tate, M. W., Bilderback, D., Fontes, E., Wixted, R., Barna, S. & Gruner, S. M. (1999). *Tests of a prototype pixel array detector for microsecond time-resolved X-ray diffraction. J. Synchrotron Rad.* **6**, 1096–1105.

Sarvestani, A., Besch, H. J., Junk, M., Meissner, W., Pavel, N., Sauer, N., Stiehler, R., Walenta, A. H. & Menk, R. H. (1998). *Gas amplifying hole structures with resistive position encoding: a new concept for a high rate imaging pixel detector*. *Nucl. Instrum. Methods Phys. Res. A*, **419**, 444–451.

International Tables for Crystallography (2012). Vol. F, Chapter 7.2, pp. 183–188.

Chapter 7.2. CCD detectors

M. W. Tate, E. F. Eikenberry and S. M. Gruner

7.2.1. Overview

After more than 20 years of refinement, CCD (charge-coupled device) detectors have emerged as the most useable and accurate large-area detectors available for the X-ray energies of interest to crystallographers. CCDs are familiar as the imagers in television and digital cameras, but the scientific grade devices used in detectors are larger and have more pixels and a lower noise amplifier. CCD detectors are an assembly of several components: an energy converter (*e.g.* phosphor), an optical relay with or without gain (fibre optics, lenses and/or intensifier) and the imaging CCD.

Although many configurations have been used in the past, improvements in the size and quality of fibre-optic tapers have led to the possibility of direct coupling – eliminating intensifiers and lenses – so long as other components are carefully optimized at the same time (Eikenberry *et al.*, 1991). Optimizations include the phosphor, the CCD and electronics, and the elimination of unneeded optical interfaces. Current commercial designs employ just three elements: phosphor, taper and CCD (Fig. 7.2.1.1). This concept enabled the use of large tapers, machined square at the front, that can be stacked together to form mosaic arrays. Consequently, there is now no inherent limit to the size of a CCD detector.

7.2.2. CCD detector assembly

Any practical detector requires compromises in the choice of components to optimize those aspects most important to the diffraction problem at hand. The optimization of one detector characteristic often adversely affects other characteristics, so that it is often difficult to identify the 'best' component. Considerations are given below to aid in making judicious choices.

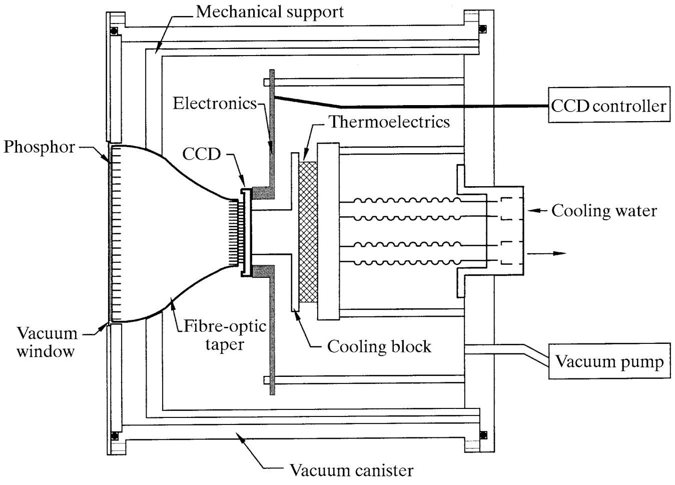


Figure 7.2.1.1
Schematic of a single-module CCD detector. The thin phosphor screen is behind a light- and vacuum-tight vacuum window and is coupled to a fibre-optic taper, which is, in turn, coupled to a CCD. The CCD is thermoelectrically cooled to 213 K and housed in a vacuum cryostat. Reproduced with permission from Tate *et al.* (1995).

Most fundamentally, a viable X-ray detector must have good quantum efficiency (see Section 7.1.1 in Chapter 7.1). Necessary, but not sufficient, are a high stopping power for X-rays and a large average signal per X-ray recorded in the CCD. The input signal passes through a sequence of stages and is transformed several times in the process. The statistics of this process are governed by the quantized nature of the signal, whether it be the initial X-ray photon, the visible photons produced in the phosphor, intermediate photoelectrons in intensifiers or the integrated charge in the CCD. To maintain a high detective quantum efficiency (DQE), the associated number of quanta per X-ray must be kept well above unity at each stage in the 'quantum chain'. There are several approaches that meet this criterion.

Crystallography applications generally benefit from large detective areas, whereas typical scientific CCDs are quite modest in size (*circa* 25 × 25 mm). The usual solution is to use fibre-optic tapers (which are more efficient demagnifiers than lenses; see Deckman & Gruner, 1986) to optically reduce a diffraction image excited in a larger phosphor screen. However, since optical image reduction is inherently inefficient, the reduction ratio is usually limited to about 4:1 before the number of visible photons per X-ray transmitted to the CCD becomes unacceptably low. Higher reduction ratios require image intensification *before* reduction (Moy, 1994; Naday *et al.*, 1995; Tate *et al.*, 1997) or there will be an unacceptable loss in DQE. Intensification *after* reduction can result in the same average recorded signal per X-ray. However, there is a significant probability that no quanta at all make it through the chain for many of the incident X-rays, thereby lowering the number of X-rays actually 'counted'.

Properties of the individual components affect other important detector characteristics as well. Below is a summary of important parameters for each of the components. Performance variations of CCD detectors from different vendors can most often be traced to the quality of the phosphor screen and the calibrations that are applied to the detector.

Phosphors. Although there are a bewildering variety of phosphor types, only a few are typically used with X-ray detectors (Shepherd *et al.*, 1995). A dense, high atomic number material is necessary to make the thin screens required for good spatial resolution while maintaining high X-ray stopping power. Gd_2O_2S:Eu offers high light output (>200 visible photons per 8 keV X-ray) and an emission spectrum matched to the typical CCD's spectral sensitivity peak in the red. Although there is a fairly prompt emission of most of the light (<1 ms to 10%), there is a long-term persistence which decays according to a power law: bright spots glow for seconds after an exposure has ended. This severely limits the dynamic range during fast framing, such as might be encountered in a synchrotron environment.

Other dopants for Gd_2O_2S, such as Tb and Pr, have much shorter persistence and are better suited to higher frame rates. These phosphors yield somewhat lower CCD signals because they emit fewer visible photons per X-ray and because their blue–green emission is less well matched to the CCD spectral sensitivity. Interestingly, Gd_2O_2S:Tb has one of the slowest 'prompt' emission (exponential decay) time constants, but is one of the fastest phosphors to decay to 10^{-4} (<10 ms), resulting in low persistence.

Thicker phosphors are needed at higher X-ray energies (>15 keV) to maintain high stopping power. This generally reduces spatial resolution. However, structured phosphors offer a way to increase thickness while limiting the lateral spread of the light. For example, CsI:Tl can be grown as an array of columnar crystals, resulting in a screen with enhanced resolution (Stevels & Schrama-de Pauw, 1974*a*, *b*; Moy, 1998). The spectral mismatch of this green-emitting phosphor is offset by the increased signal per X-ray expected at these higher X-ray energies. Recently, a (Zn,Cd)Se phosphor has been described that has excellent characteristics for X-ray energies above 12 keV; the stopping power at lower energies is compromised by absorption edges (Bruker AXS Inc.).

Fibre optics and lenses. Light transmission in image reduction is limited by n.a. × M^2, where n.a. is the numerical aperture of the optical system and M is the linear magnification factor ($M < 1$ corresponds to reduction). Lens systems have low values of n.a. when used in reduction, so there is only 2% light transmission for a 3:1 reduction with an f/1.0 lens. With a much higher n.a., fibre optics can typically transmit 13% with the same 3:1 reduction (Coleman, 1985). Such reducing tapers are produced by locally heating, then pulling, a fused bundle of optical fibres. Each fibre within the bundle becomes tapered in this process, thereby reducing the image scale from front to back of the bundle. The bundle structure introduces a characteristic 'chicken-wire' pattern into the image which must be removed *via* intensity calibration (Section 7.2.3). Tapers up to 165 mm in diameter are available.

To obtain good resolution, extramural absorbing fibres (EMAs) must be placed in the fibre optic to absorb light that propagates between fibres. These EMAs are often a more effective absorber in the blue part of the spectrum, with the result that a red-emitting phosphor yields somewhat poorer resolution. Another concern with fibre optics is radioactivity in the glass used to produce the optic, which manifests itself as random flashes ('zingers') in both the phosphor and the CCD (Section 7.2.3).

Although fibre-optic coupling is preferred in many situations, lenses are appropriate for use with image intensification, or in cases of image magnification, such as for microtomography, using high-resolution screens.

Image intensifiers allow large demagnification ratios without undue DQE loss (Moy, 1994). In these vacuum-tube devices, visible light produces photoelectrons in the photocathode of the intensifier, which in turn are accelerated under high potential onto a secondary phosphor screen. Each photoelectron excites many photons in the secondary phosphor, giving light amplification. Image resolution is retained either by magnetic, electrostatic or proximity focusing of the electrons. In the case of microchannel plate intensifiers, photoelectrons are restricted to cascade down hollow fibres. Intensifiers often have problems with stability and linearity, and they degrade resolution. Phosphor afterglow in the secondary phosphors is also a consideration. Intensifiers typically limit the input format to less than 80 mm in diameter, although one design uses a large (230 mm) radiographic intensifier (Moy, 1994; Hammersley *et al.*, 1997). The greatest drawbacks of image intensifiers are cost, availability, susceptibility to magnetic fields and lack of robustness.

CCDs. The low noise and high sensitivity of scientific CCDs have made CCDs the preferred imaging device in X-ray applications. Visible photons are converted to charge carriers in the silicon of the CCD, with a 30–40% quantum efficiency (q.e.) in the red, dropping to 5% in the blue owing to increased absorption by the controlling gate structure (the q.e. can be considerably higher for back-illuminated devices which, however, are expensive, fragile and difficult to obtain). The 'read noise' (noise with zero charge in the pixel) is routinely less than 10 electrons for scientific grade chips with pixel read rates of less than 1 MHz. When coupled to a good phosphor screen with a modest (*circa* 3:1) fibre-optic taper, signal levels in the CCD are typically 10–30 recorded electrons per 10 keV X-ray. This is above the CCD read-noise level, thereby allowing low-dose quantum-limited X-ray imaging.

A physically large CCD chip is usually desired to maximize the detector area, given the limited image reduction ratio. Devices from 2–4 cm on an edge with 1000 × 1000 to 2000 × 2000 pixels are available. These sizes match well to the largest of the fibre-optic tapers and are therefore most often used in detectors. Larger-format CCDs are available, but are harder to obtain and are expensive.

CCDs are usually cooled well below room temperature to minimize thermally generated dark current, an unwanted source of noise. The dark current drops by a factor of 2 for every 5–7 K drop in temperature. At 233 K, there is roughly one electron per pixel per second dark current for normal clocking operation. Since the dark current changes with temperature, the temperature must be well regulated (to within ±0.1 K to subtract the dark charge from an image reproducibly.

Most of the dark current comes from surface defects, not from the bulk silicon. Multiphase pinned (MPP) CCDs use charge implants within the pixel structure to move the charge collection region away from the surface, thereby reducing the dark current by several orders of magnitude. This is usually accompanied by a significant reduction in the maximum charge capacity ('full well') of a pixel. Even so, MPP CCDs are becoming the norm in most X-ray detectors.

Directly exposed CCDs. CCDs can directly image X-rays, although typical CCDs are not very efficient, since the charge collection region in the silicon is very thin (<10 μm). Thicker depletion regions can be fabricated in high-resistivity silicon wafers, improving X-ray collection efficiency to greater than 30% at 8 keV. Direct conversion of X-rays in silicon produces a large signal with excellent energy resolution. It is therefore possible to use the chip as an X-ray counting detector with the ability to discriminate in energy. To retain X-ray-energy measurement capability, however, there must be less than one X-ray per pixel per frame to avoid signal overlap. This requires a fast readout as well as large amounts of disk storage to handle the large number of files. The X-rays damage the CCD's electronic structure, resulting in a higher dark current within the exposed pixels. Care must be taken to shield non-imaging parts of the CCD, such as the output amplifier, which will adversely affect chip performance at even lower radiation dose.

7.2.3. Calibration and correction

Inhomogeneities within the detector components introduce nonuniformities in the output image of several per cent or more, both as geometric distortion and as nonuniformity of response. The response of the system varies not only with position, but also with the angle of incidence and X-ray energy. Optimal calibration of the detector should take into account the parameters of the X-ray experiment, seeking to mimic the experimental conditions as closely as possible: a uniform source of X-rays of the proper energy positioned in place of the diffracting crystal would be ideal. Realizing such a source is somewhat problematic, so the calibration procedure is often broken down into several independent steps. Calibration procedures are detailed in Barna *et al.* (1999) and are summarized below.

7.2.3.1. *Dark-current subtraction*

It is important to remove both the electronic offset and the accumulated dark charge from an X-ray image. Since the integrated dark current varies from pixel to pixel and with time, a set of images needs to be taken (with no X-rays), matched in integration time to the X-ray exposures. With a properly temperature-stabilized detector, the background images may be acquired in advance and used throughout an experiment. Because the background image has noise, it is common to average a number of separate backgrounds to minimize the noise.

7.2.3.2. *Removal of radioactive decay events*

Cosmic rays and radioactive decay of actinides in the fibre-optic glass produce large-amplitude isolated signals ('zingers') within an X-ray image. These accumulate randomly in position and in time. For the short exposures typical at synchrotron sources and for data sets with highly redundant information, the few diffraction spots affected by these events can be discarded with statistical analysis. For longer exposures, or for data with less redundancy, two (or more) nominally identical exposures can be taken and then compared pixel by pixel to remove spurious events (see Barna *et al.*, 1999).

7.2.3.3. *Geometric distortion*

Geometric distortions arise in the optical coupling of the system. If they are stable, the distortions can be mapped and corrected. Long-term stability is possible for a phosphor fibre-optically coupled directly to a CCD, since all distortions are mechanically fixed. Intensifier-based systems are subject to changes in magnetic and electric fields, hence stability is more of a problem.

Geometric distortions may be either continuous or discontinuous. Fibre optics often have shear between adjacent bundles of fibres. In this defect, one group of fibres will not run parallel to a neighbouring group, causing a discontinuity in the image. Rather than dealing with such discontinuities, tapers with low shear (less than one pixel maximum) are usually specified. Even with low shear, there is a continuous distortion (several per cent), which varies slowly over the face of the detector. Such distortion is inevitable, as the temperature profile cannot be precisely controlled in the large block of glass comprising the fibre optic as it is processed. To map this distortion, an image is taken of a regular array of spots. Such an image can be made by illuminating a shadow mask of equally spaced holes with a flood field of X-rays. Holes 75 μm in diameter spaced on a 1 × 1 mm square grid are adequate for mapping the distortions present in most fibre-optic tapers. Such masks have been lithographically fabricated in an X-ray opaque material, such as 50 μm tungsten foil (Barna *et al.*, 1999).

Given an image produced with this X-ray mask, the displacement map for every pixel in the original image can be computed as follows: find the centroid of each mask spot and its displacement relative to an ideal lattice. The array of spot positions and associated displacements can then be interpolated to find the displacement for each pixel in the original distorted image.

The displacement of a pixel from original to corrected image will not in general be a whole number. Rather, the intensity of a pixel will be distributed in a local neighbourhood of pixels in the corrected image centred about the position given in the displacement map. The size of the neighbourhood depends on the local dilation or contraction of the image; typically the intensity will be distributed in one to nine pixels. This distribution procedure yields a smooth intensity mapping. Applying corrections to mask images that have been arbitrarily displaced shows that the distortion correction algorithm is good to better than 0.25 pixels (Barna *et al.*, 1999).

The geometrical distortion is tied intimately to the correction of the response of the system (see below). Since distortions produce local regions of dilation or contraction of the image, pixels will, in general, correspond to varying sizes. This variation must be included in the flat-field calibration.

7.2.3.4. *Flat-field corrections*

Variation in response arises from nonuniformity in the phosphor, defects in the fibre optics and pixel-to-pixel variation in the sensitivity of the CCD itself. Of these, variations in the fibre optics are usually most pronounced, often with decreased transmission at the bundle interfaces, resulting in the characteristic 'chicken-wire' pattern. The response of the system is generally stable with time, although exposure to the direct beam at synchrotron sources will cause colour centres to form in the glass of the fibre optic. Recalibration of the system will correct for the reduced transmission in the exposed spot.

Light output from the phosphor depends on a number of factors: phosphor thickness, X-ray energy, angle of incidence and depth of conversion within the screen. X-ray photons are absorbed within phosphor grains with an exponentially decaying distribution as they travel deeper into the phosphor layer. Making the phosphor layer thicker increases the probability that the X-ray photons are absorbed, yielding an increase in the signal produced. But thicker phosphor layers also scatter the visible photons which one desires to collect in the CCD. At some point, the loss of light will be greater than the increase in X-ray stopping power and the net response will decrease. For one particular phosphor preparation, this appears to happen at *circa* 85% stopping power. The phosphor-screen resolution also falls with increasing screen thickness, although surprisingly slowly owing to the diffusive nature of the light scatter. These effects are discussed in Gruner *et al.* (1993).

Consider a given phosphor layer with thickness nonuniformity. For X-ray energies where the stopping power of the phosphor is low, regions of the phosphor with a thickness larger than the mean will be brighter owing to the increased X-ray stopping power. For energies at which X-rays are strongly absorbed, the increased opacity of the thicker regions of the phosphor will cause the response to go down. This illustrates the importance of calibrating the response of the system to the particular X-ray energy of interest.

In like manner, X-rays impinging on the phosphor at an angle away from the normal are presented with a longer path length and hence an increased stopping power. Also, because of the oblique angle, the distribution of visible-light production will be shifted toward the phosphor surface. Again, for strongly absorbed X-ray energies, the increase in the optical path for the light will cause the recorded signal to fall, whereas for X-rays not as strongly absorbed, the signal will increase.

To map the nonuniformity in response, one would ideally use a uniform source of X-rays of the proper energy placed at the position of the sample. This would calibrate the detector with the proper energy and angle of incidence for the diffraction data to be corrected. Correction factors are computed from a series of images taken of this uniform source. Sufficient numbers of X-rays per pixel must be collected to reduce the shot noise in the X-ray measurement to the required level (*e.g.* 40 000 X-rays per pixel must be acquired to correct to 0.5%).

Providing a truly uniform source with an arbitrary X-ray energy and angular distribution is difficult at best. Other sources can be used, however, with good results. Amorphous samples containing a variety of elements can be fabricated which produce X-ray fluorescence at various wavelengths when excited by a synchrotron beam (Moy *et al.*, 1996). These can be placed at the position of the sample, thereby mimicking the angular distribution of X-rays from the experiment. The fluorescence is not uniform in space, however, so that the actual distribution must be mapped by some means. Once mapped, however, these samples provide a stable calibration source.

Another alternative is to separate the calibration procedure into several parts, mapping the dependence at normal incidence and treating the angular dependence as a higher-order correction. By moving an X-ray source sufficiently far away, the detector can be illuminated at near-normal incidence with excellent uniformity. For example, an X-ray tube at 1 m distance can produce a field with uniformity better than 0.5% over a 10 × 10 cm area. A sum of images with sufficient X-ray statistics, taken with the area dilation of each pixel computed through the geometric distortion calibration, can be used to compute a pixel-by-pixel normalization factor. Again, this should be determined for the X-ray energy of interest. In practice, the appropriate energy may be approximated by a linear combination of several energies. However, the proper coefficients will need to be determined empirically. This can be judged in diffraction having broad diffuse features, because imperfections in the phosphor stand out clearly when the data are corrected using factors derived from the wrong X-ray energy.

7.2.3.5. *Obliquity correction*

It has been found empirically that the light output for a given X-ray energy has a quadratic variation with the angle of incidence. The quadratic coefficient varies with X-ray energy and may be either positive or negative. The angular dependence may be measured by illuminating the detector with a small stable spot of X-rays and recording the integrated dose at various angles of incidence. Given the placement of the detector in relation to the beam and sample, the angle of incidence at a particular pixel can be computed and can be used to find the correction factor needed. With this method, a change in the experimental setup does not require a new calibration, just the computation of a new set of coefficients. The combination of energy and obliquity sensitivity varies slowly and may be approximated by a quadratic or cubic fit to a surface as a function of X-ray energy and angle. The few coefficients defining this surface allow quick computation of the combined energy and obliquity factor with which to multiply the local flat-field correction for X-rays of known incident energy and angle.

The obliquity correction is often ignored, since the solution of structures from X-ray diffraction typically includes a temperature factor which also varies with angular position. Uncorrected angular dependence of detector response will be convoluted with the true temperature factor and often does not impede the solution of the structure.

These procedures do not allow correction to arbitrary accuracy, however. The calibration data are taken with uniform illumination, whereas diffraction spots are localized. Given a nonzero point spread in the detector, the computed correction factor arises from a weighted average of many illuminated pixels. The signal from a diffraction spot only illuminates a few pixels, so the true factor might well be different. This should be less of a problem as diffraction spots become larger, becoming more like a uniform illumination. Measurement for one detector showed 75 µm spots could be measured to 1% accuracy, whereas 300 µm spots could be measured to 0.3% accuracy (Tate *et al.*, 1995).

7.2.3.6. *Modular images*

The size of available fibre optics and CCDs and the inefficiencies of image reduction limit the practical imaging area of a single CCD system. Closely stacked fibre-optic taper CCD modules can be used to cover a larger area. Although the image recorded from each module could be treated as independent in the analysis of the X-ray data, merging the sub-images into one seamless image facilitates data processing. Each module will have its own distortion and intensity calibration. It is no longer possible to choose an arbitrary lattice onto which each distorted image will be mapped: the displacement and scaling must be consistent between the modules. This would be accomplished most easily by having a distortion mask large enough to calibrate all modules together, although it is possible to map the intermodule spacing with a series of mask displacements.

Flat-field correction proceeds as in the case of a single module detector after proper scaling of the gain of each unit is performed. There can potentially be a change in the relative scale factors between modules, since each is read though an independent amplifier chain. Multimodule systems emphasize the need for enhanced stability and ease of recalibration.

7.2.4. Detector system integration

Hardware interfacing. CCDs are operated, during both data acquisition and readout, by a dedicated hardware controller attached to a host computer, generally a PC. The requirements for the controller are complex and quite stringent in order to obtain low-noise operation. As with any high-speed electronics, noise increases with speed. Typically, pixel read rates of 100–500 kHz are used, although higher read rates can be used (and still preserve low-noise performance) for CCDs with multistage on-chip amplifiers. Some CCDs have multiple output amplifiers that increase pixel throughput by using parallel digitization channels. The entire CCD can also be read at reduced resolution by analogue summation of adjacent rows and/or columns on the chip (binning). This has the added benefit of increasing the signal-to-noise ratio for signals with low spatial frequency since there are fewer digitizations. Binning is highly recommended whenever the reduced resolution can be tolerated.

The time resolution of the detector can be further increased if the CCD array is used as frame storage. In this case, a portion of the imaging area is masked to X-rays, making it available for storage. The exposed area can be shifted rapidly into the masked area and a second exposure begun. Storage for five to ten subframes can easily be configured before readout is necessary. The time resolution is ultimately limited by the phosphor decay time and the time needed to shift the image. Although most

CCDs are capable of being operated in very flexible ways, flexible CCD controllers are expensive. The consequence is that few commercial CCD X-ray detectors permit use of all the available options.

The detector itself is contained in a cryostat with the low-noise parts of the controller nearby, either in a separate box connected by a short cable or mounted inside the cryostat itself. A longer cable carries the time-multiplexed digitized data to the computer. High-speed serial data technologies are under investigation to simplify this connection and will become imperative for the much larger format detectors that are being developed.

Several installations have constructed a safety shield in front of the detector that opens only when data are being collected. This device helps to protect the delicate front surface of the detector and is highly recommended.

Data acquisition software. There is a wide spectrum of computer configurations surrounding CCD detectors. The major tasks to be performed are operating the detector, controlling the beamline, storing raw data, correcting images and analysing diffraction patterns. In home laboratories, where exposure times are relatively long, a single PC typically handles these tasks. In another arrangement, the detector controller is really an embedded system, mostly unseen by the operator, making the detector a remote image server. The raw data or corrected images come to the user's workstation where subsequent analysis is performed. This circumvents the problem that the detector computer may be running a different operating system from the workstation. At storage-ring sources, where the data volume is very large, the detector is almost always configured as a remote image server; the user's workstation does not even need to be nearby. Clusters of remote computers that can perform tasks in parallel become attractive for streamlined data collection, correction and analysis from large data sets. Remote analysis over the internet is being explored by several storage-ring facilities.

Control software should be easy to use, but flexible and extensible. It should be easy to set up experiments and sequence the individual steps in an experiment: exposure, readout, correction, storage and crystal movement, and wavelength change for MAD experiments. Extensible software would permit a user-written macro to be run at each step in place of the detector primitive that is provided. For instance, if it were desired to collect two images at different exposure times for each position of the crystal, extensible software would make it easier to set up the experiment. Finally, the software should permit access to all of the readout modes of the detector. For instance, a detector may be capable of rapidly scanning a small region of interest for alignment purposes, or it may be capable of streak-mode operation for certain types of time-resolved experiments. Available CCD detector software for macromolecular applications has room for much improvement. Hopefully, software will continue to undergo rapid development. Standardization is especially needed.

7.2.5. Applications to macromolecular crystallography

Storage rings. CCD detectors have gained widespread acceptance for macromolecular crystallography at storage-ring sources, in part because of the high-quality data they give, but more for their speed, convenience and efficiency. Accurate data to high resolution are especially important for MAD phasing, and CCD detectors excel in this application. In the past with film, or even storage phosphors, teams of perhaps ten people were required to perform a synchrotron experiment; today, a single person per shift can perform an experiment. With increasing beam flux, improved X-ray optics and faster CCDs, it is often possible to collect full data sets in little more than an hour. Anticipated improvements in speed for CCD detectors should soon make it feasible to collect fine-sliced rotation data routinely; these data are expected to yield better structure solutions.

Home laboratories. Acceptance of CCD detectors for macromolecular crystallography at home laboratories has been slower, in part because there is not such a premium on speed, and in part because of cost. Diffracted spot sizes are larger than at synchrotrons, so highly accurate data should be obtainable. Fully automatic storage phosphor systems work quite well with conventional sources and at this time are lower in cost than large CCD detectors. However, they have a minimum cycle time, caused by the mechanics of the readout scheme, and the required exposure for a strongly diffracting crystal can best this time by a wide margin. Thus, for strongly diffracting specimens, CCD detectors can be significantly more efficient.

7.2.6. Future of CCD detectors

The basic principles of CCD detector technology are now well developed, but various incremental improvements have already been demonstrated and may be expected in commercial detectors. These include larger detector areas, faster read times (owing to both faster electronics and multiamplifier CCDs), more flexible control electronics, better optimized phosphors and calibrations, and, especially, better software. Lower-cost CCD detectors would certainly be welcome. It is easily predicted that the application of CCD detectors will continue to increase rapidly for at least several more years until displaced by even better technologies, such as pixel array detectors (see Section 7.1.4 in Chapter 7.1).

References

Barna, S. L., Tate, M. W., Gruner, S. M. & Eikenberry, E. F. (1999). *Calibration procedures for charge-coupled device X-ray detectors. Rev. Sci. Instrum.* **70**, 2927–2934.

Coleman, C. I. (1985). *Imaging characteristics of rigid coherent fiber optic tapers. Adv. Electron. Electron Phys.* **64B**, 649–661.

Deckman, H. W. & Gruner, S. M. (1986). *Format alterations in CCD based electro-optic X-ray detectors. Nucl. Instrum. Methods Phys. Res. A*, **246**, 527–533.

Eikenberry, E. F., Tate, M. W., Belmonte, A. L., Lowrance, J. L., Bilderback, D. & Gruner, S. M. (1991). *A direct-coupled detector for synchrotron X-radiation using a large format CCD imaging array. IEEE Trans. Nucl. Sci.* **38**, 110–118.

Gruner, S. M., Barna, S. L., Wall, M. E., Tate, M. W. & Eikenberry, E. F. (1993). *Characterization of polycrystalline phosphors for area X-ray detectors. Proc. SPIE*, **2009**, 98–108.

Hammersley, A. P., Brown, K., Burmeister, W., Claustre, L., Gonzalez, A., McSweeney, S., Mitchell, E., Moy, J.-P., Svensson, S. O. & Thompson, A. W. (1997). *Calibration and application of an X-ray image intensifier/charge-coupled device detector for monochromatic macromolecular crystallography. J. Synchrotron Rad.* **4**, 67–77.

Moy, J.-P. (1994). *A 200 mm input field 5–80 keV detector based on an X-ray image intensifier and CCD camera. Nucl. Instrum. Methods Phys. Res. A*, **348**, 641–644.

Moy, J.-P. (1998). *Image quality of scintillator based X-ray electronic imagers. Proc. SPIE*, **3336**, 187–194.

Moy, J. P., Hammersley, A. P., Svensson, S. O., Thompson, A., Brown, K., Claustre, L., Gonzalez, A. & McSweeney, S. (1996). *A novel technique for accurate intensity calibration of area X-ray detectors at almost arbitrary energy. J. Synchrotron Rad.* **3**, 1–5.

Naday, I., Ross, S., Kanyo, M., Westbrook, M., Westbrook, E. M., Phillips, W. C., Stanton, M. J. & O'Mara, D. (1995). *The gold detector: modular CCD area detector for macromolecular crystallography*. *Proc. SPIE*, **2415**, 236–249.

Shepherd, J. A., Gruner, S. M., Tate, M. W. & Tecotzky, M. (1995). *A study of persistence in gadolinium oxysulfide X-ray phosphors*. *Proc. SPIE*, **2519**, 24–30.

Stevels, A. & Schrama-de Pauw, A. (1974*a*). *Vapour-deposited CsI:Na layers. I. Morphologic and crystallographic properties*. *Philips Res. Rep.* **29**, 340–352.

Stevels, A. & Schrama-de Pauw, A. (1974*b*). *Vapour-deposited CsI:Na layers. II. Screens for application in X-ray imaging devices*. *Philips Res. Rep.* **29**, 353–362.

Tate, M. W., Eikenberry, E. F., Barna, S. L., Wall, M. E., Lowrance, J. L. & Gruner, S. M. (1995). *A large-format high-resolution area X-ray detector based on a fiber-optically bonded charge-coupled device (CCD)*. *J. Appl. Cryst.* **28**, 196–205.

Tate, M. W., Eikenberry, E. F. & Gruner, S. M. (1997). *Coupling format variations in X-ray detectors based on charge-coupled devices*. *Rev. Sci. Instrum.* **68**, 47–54.

International Tables for Crystallography (2012). Vol. F, Chapter 8.1, pp. 189–204.

PART 8. SYNCHROTRON CRYSTALLOGRAPHY

Chapter 8.1. Synchrotron-radiation instrumentation, methods and scientific utilization

J. R. Helliwell

8.1.1. Introduction

Synchrotron radiation (SR) has had a profound impact on the field of protein crystallography. The properties of high spectral brightness and tunability have enabled higher-resolution structure determinations, multiple-wavelength anomalous-dispersion (MAD) techniques, studies of much larger molecular weight structures, the use of small crystals and dynamical time-resolved structural studies. The use of SR required development of suitable X-ray beamline optics for focusing and monochromatization of the beam, which had to be stable in position and spectral character, for rotating-crystal data collection. Finely focused polychromatic beams have been used for ultra-fast data collection with the most advanced SR sources, where a single bunch pulse of X-rays can be strong enough to yield Laue diffraction data. The optimal recording of the diffraction patterns has necessitated the development of improved area detectors, along with associated data-acquisition hardware and data-processing algorithms. Sample cooling and freezing have reduced and greatly diminished, although not eliminated, radiation damage, respectively. In turn, even smaller crystals have been used. However, new X-radiation damage challenges are being reached. The low emittance of SR sources, with their small source size and beam divergence, corresponds well with the small size and low mosaicity of protein crystal samples. The evolution of SR source spectral brightness each year over the last twenty years has changed by many orders of magnitude, a remarkable trend in technical capability.

8.1.2. The physics of SR

The physics of the SR source spectral emission was predicted by Iwanenko & Pomeranchuk (1944) and Blewett (1946), and was fully described by Schwinger (1949). It is 'universal' to all machines of this type, *i.e.*, wherever charged particles such as electrons (or positrons) travel in a curved orbit under the influence of a magnetic field, and are therefore subject to centripetal acceleration. At a speed very near the speed of light, the relativistic particle emission is concentrated into a tight, forward radiation cone angle. There is a continuum of Doppler-shifted frequencies from the orbital frequency up to a cutoff. The radiation is also essentially plane-polarized in the orbit plane. However, in high-energy physics machines, the beam used in target or colliding-beam experiments would be somewhat unstable; thus, while pioneering experiments ensued through the 1970s, a considerable appetite was stimulated for machines dedicated to SR with stable source position, for fine focusing onto small samples such as crystals and single fibres, and with a long beam lifetime for more challenging data collection. Crystallography has been both an instigator and major beneficiary of these developments through the 1970s and 1980s onwards. An example of a machine lattice (the ESRF) is shown in Fig. 8.1.2.1.

The properties of synchrotron radiation can be described in terms of the well defined quantities of high flux (a large number of photons), high angular brightness (also well collimated), high spectral brightness (also a small source size and well collimated), tunable, polarized, defined time structure (fine time resolution) and exactly calculable spectra. The precise definition of the spectral brightness is

$$\text{Spectral brightness} = \text{photons per s per mm}^2 \text{ per mrad}^2 \text{ per } (0.1\% \text{ bandwidth}). \qquad (8.1.2.1)$$

Care needs to be exercised to check precisely the definition in use (Mills *et al.*, 2005). The mrad^2 term refers to the radiation solid angle delivered from the source, and the mm^2 term to the source cross-sectional area. Mills *et al.* (2005) concluded that the units given in equation (8.1.2.1), which do not follow the SI code for

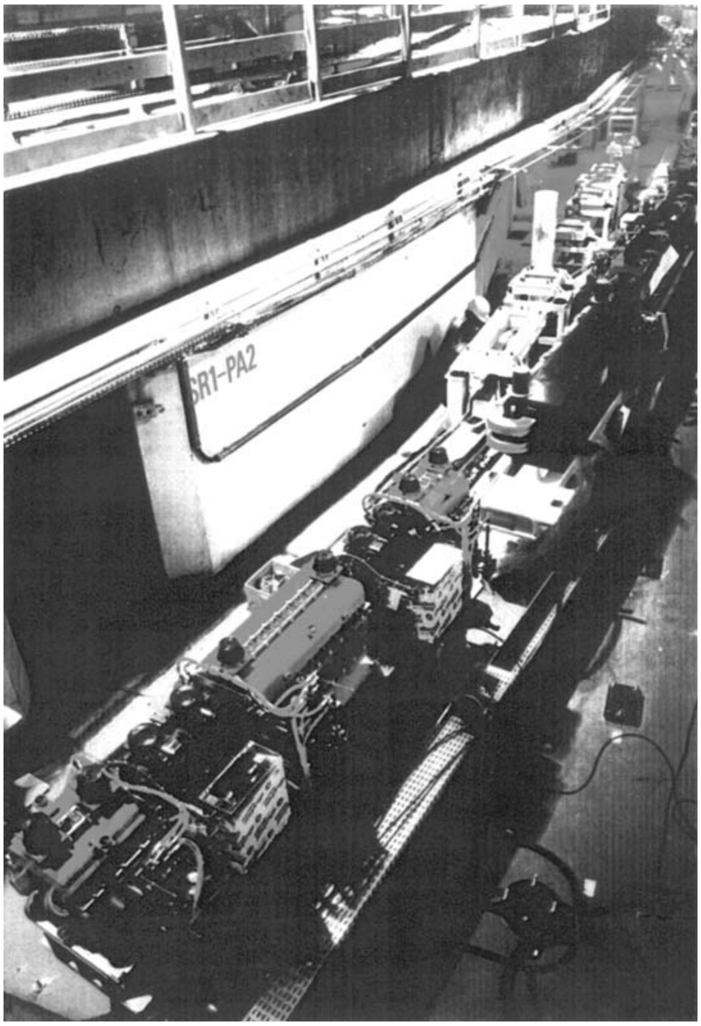


Figure 8.1.2.1
The ring tunnel and part of the machine lattice at the ESRF, Grenoble, France.

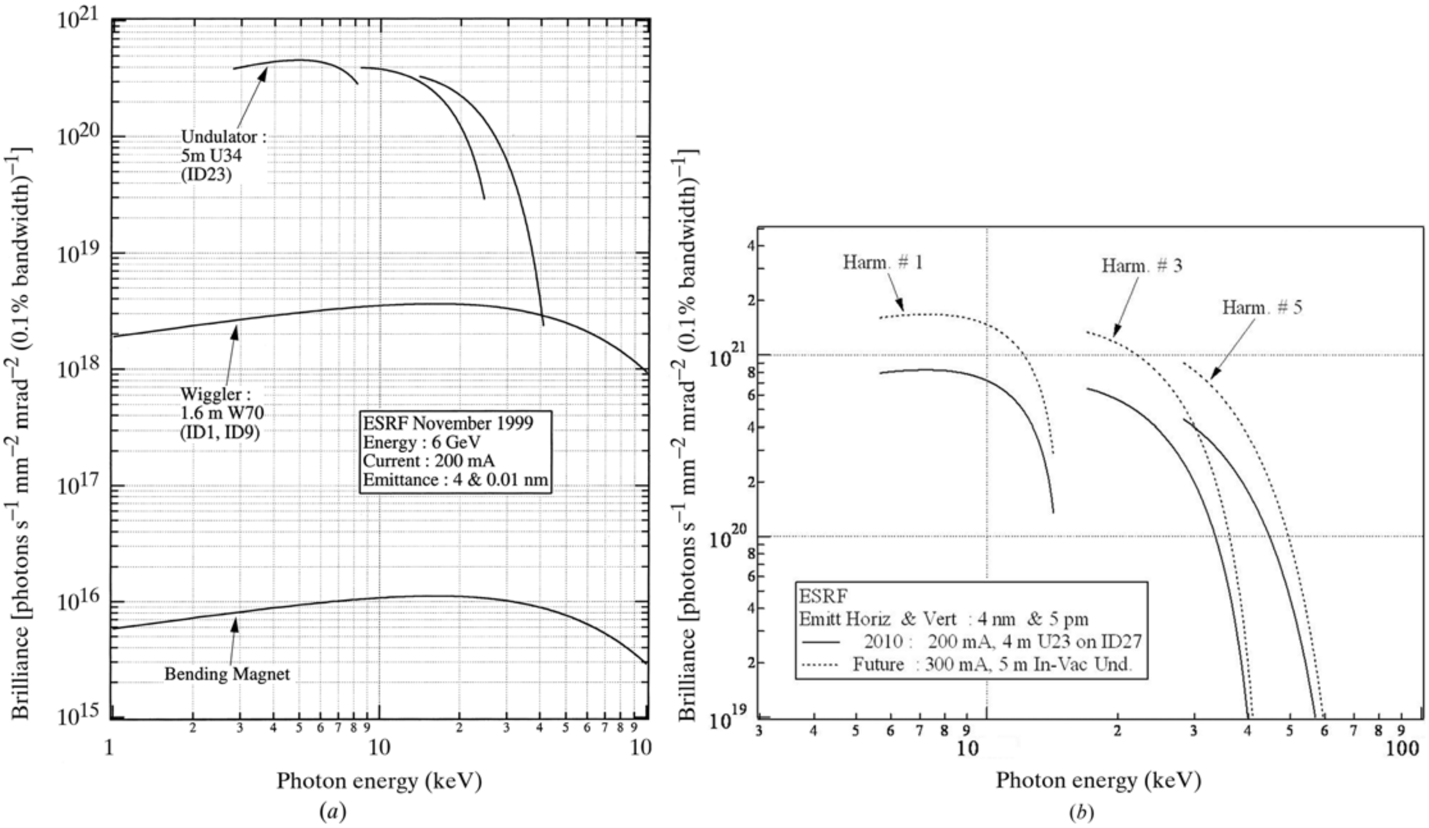


Figure 8.1.2.2

SR spectra. (*a*) Spectral brightness (also referred to as brilliance) of different SR source types (undulator, multipole wiggler and bending magnet) as exemplified by such types of sources at the ESRF. For the undulator, the tuning range (*i.e.* as the magnet gap is changed) is indicated. (*b*) Brilliance produced by the in-vacuum undulator of cell 27 of the ESRF dedicated to high-pressure studies. The plain curve corresponds to the condition in use as of September 2010. Further increase in brilliance (dotted curve) is expected in the years to come by increasing the ring current, increasing the length of the undulator and further decreasing the vertical emittance. Kindly provided by Dr Pascal Elleaume, ESRF, Grenoble, France.

units, are so ensconced in the field 'that a drive to change this would only lead to more confusion rather than more clarity in the descriptions of synchrotron-radiation sources'. At the ESRF the term brilliance is firmly ensconced in house and with its large user community, and so is the label for the *y* axes used in Fig 8.1.2.2.

Another useful term is the machine emittance, ε. This is an invariant for a given machine lattice and electron/positron machine energy. It is the product of the divergence angle, σ', and the source size, σ:

$$\varepsilon = \sigma\sigma'. \qquad (8.1.2.2)$$

The horizontal and vertical emittances need to be considered separately.

The total radiated power, Q (kW), is expressed in terms of the machine energy, E (GeV), the radius of curvature of the orbiting electron/positron beam, ρ (m), and the circulating current, I (A), as

$$Q = 88.47E^4I/\rho. \qquad (8.1.2.3)$$

The opening half-angle of the synchrotron radiation is $1/\gamma$ and is determined by the electron rest energy, mc^2, and the machine energy, E:

$$\gamma^{-1} = mc^2/E. \qquad (8.1.2.4)$$

The basic spectral distribution is characterized by the universal curve of synchrotron radiation, which is the number of photons per s per A per GeV per horizontal opening in mrad per 1% $\delta\lambda/\lambda$ integrated over the vertical opening angle, plotted versus λ/λ_c. Here the critical wavelength, λ_c (Å), is given by

$$\lambda_c = 5.59\rho/E^3, \qquad (8.1.2.5)$$

again with ρ in m and E in GeV. Examples of SR spectral curves are shown in Fig. 8.1.2.2(*a*). The peak photon flux occurs close to λ_c, the useful flux extends to about $\lambda_c/10$, and exactly half of the total power radiated is above the critical wavelength and half is below this value.

In the plane of the orbit, the beam is essentially 100% plane polarized. This is what one would expect if the electron orbit was visualized edge-on. Away from the plane of the orbit there is a significant (several per cent) perpendicular component of polarization. Schiltz & Bricogne (2009) advocated definitions to use in the analysis of polarization-dependent phenomena that are instrument-independent and completely general. They have implemented these methods in the macromolecular phasing program *SHARP* for exploiting the polarization anisotropy of anomalous scattering in protein crystals.

8.1.3. Insertion devices (IDs)

These are multipole magnet devices placed (inserted) in straight sections of the synchrotron or storage ring. They can be designed to enhance specific characteristics of the SR, namely

(1) to extend the spectral range to shorter wavelengths (wavelength shifter);
(2) to increase the available intensity (multipole wiggler);
(3) to increase the spectral brightness *via* interference and also yield a quasi-monochromatic beam (undulator) (Fig. 8.1.2.2*b* shows the distinctly different emission from an undulator);

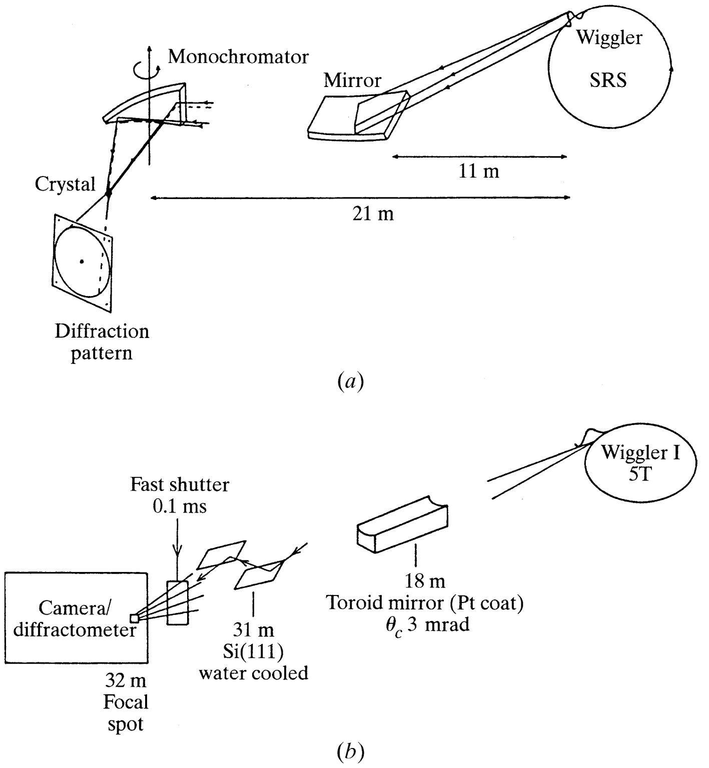


Figure 8.1.4.1
Common beamline optics modes. (*a*) Horizontally focusing cylindrical monochromator and vertical focusing mirror [shown here for station 9.6 as existed at the SRS (adapted from Helliwell *et al.*, 1986)]. (*b*) Rapidly tunable double-crystal monochromator and point-focusing toroid mirror [shown here as existed for station 9.5 at the SRS (adapted from Brammer *et al.*, 1988)].

(4) to provide a different polarization (*e.g.* to rotate the plane of polarization, to produce circularly polarized light *etc.*).

The classification of a periodic magnet ID as a wiggler or undulator is based on whether the angular deflection, δ, of the electron beam is small enough to allow radiation emitted from one pole to interfere directly with that from the next pole. In a wiggler, $\delta \gg \gamma^{-1}$, so the interference is negligible and the spectral emission (Fig. 8.1.2.2*a*) is very similar in shape to, but scaled up from, the universal curve (*i.e.* bending magnet spectral shape). In an undulator $\delta \leq \gamma^{-1}$ and the interference effects are highly significant (Fig. 8.1.2.2*b*). If the period of the ID is λ_u (cm), then the wavelengths λ_i (i integer) emitted are given by

$$\lambda_i = \frac{\lambda_u}{i2\gamma^2}\left(1 + \frac{K^2}{2} + \gamma^2\theta^2\right), \qquad (8.1.3.1)$$

where $K = \gamma\delta$.

The spectral width of each peak is

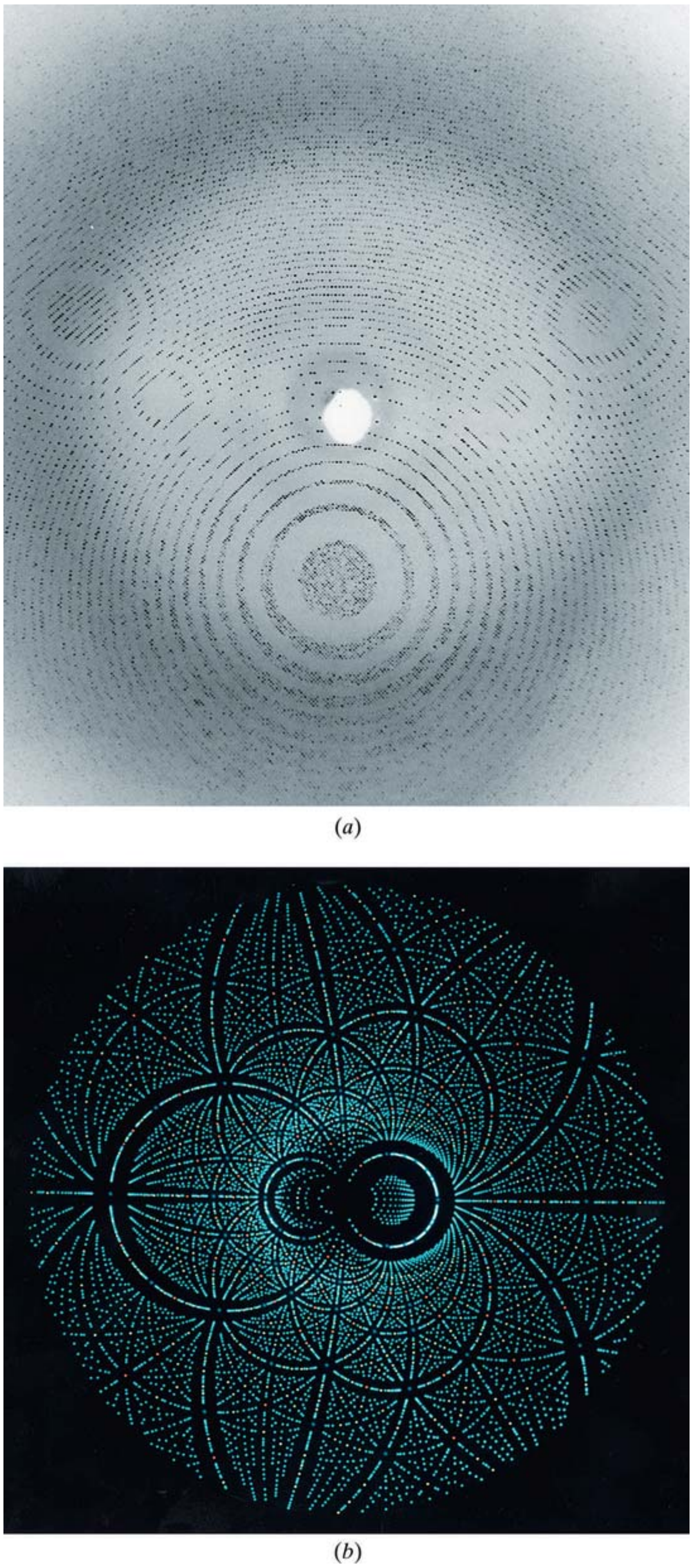
$$\Delta_i \simeq 1/iN, \qquad (8.1.3.2)$$

where N is the number of poles. The angular deflection, δ, is changed by opening or closing the gap between the pole pieces.

8.1.4. Beam characteristics delivered at the crystal sample

The sample acceptance, α [equation (8.1.4.1)], is a quantity to which the synchrotron machine emittance [equation (8.1.2.2)] should be matched, *i.e.*,

$$\alpha = x\eta, \qquad (8.1.4.1)$$


(*a*)

(*b*)

Figure 8.1.4.2
Single-crystal SR diffraction patterns. (*a*) Rhinovirus monochromatic oscillation photograph recorded at CHESS (Arnold *et al.* 1987; see also Rossmann & Erickson, 1983). (*b*) Prediction of a protein crystal Laue diffraction pattern (for an illuminating bandpass, without monochromator, $\sim 0.4 < \lambda < 2.6$ Å). The colour coding is according to the multiplicity of each spot: turquoise for singlet reflections, yellow for doublets, orange for triplets and blue for quartet or higher-multiplicity Laue spots (Cruickshank *et al.*, 1991).

where x is the sample size and η the mosaic spread. For example, if $x = 0.1$ mm and $\eta = 1$ mrad (0.057°), then $\alpha = 10^{-7}$ m rad or 100 nm rad.

Table 8.1.4.1
Internet addresses of SR facilities with macromolecular crystallography beamlines

Synchrotron-radiation source	Location	Address
ALS, Advanced Light Source	Lawrence Berkeley Lab., Berkeley, California, USA	http://www-als.lbl.gov/als/
ANKA Synchrotron	Karlsruhe, Germany	http://ankaweb.fzk.de/
APS, Advanced Photon Source	Argonne National Lab., Chicago, Illinois, USA	http://epics.aps.anl.gov/
Australian Synchrotron	Clayton, Victoria, Australia	http://www.synchrotron.org.au/
BESSY	Berlin, Germany	http://www.bessy.de/
Brazilian Synchrotron Light Laboratory	Campinas, Brazil	http://www.lnls.br/
BSRF, Beijing Synchrotron Radiation Facility	Beijing, China	http://www.ihep.ac.cn/bsrf/english/main/main.htm
CAMD, Center for Advanced Microstructures and Devices	Baton Rouge, Louisiana, USA	http://www.camd.lsu.edu/
Canadian Light Source	Saskatoon, Canada	http://www.lightsource.ca/
CHESS, Cornell High Energy Synchrotron Source	Ithaca, New York, USA	http://www.chess.cornell.edu/
Diamond Light Source	Harwell Science and Innovation Campus, Didcot, England	http://www.diamond.ac.uk/
Elettra	Trieste, Italy	http://www.elettra.trieste.it
ESRF, European Synchrotron Radiation Facility	Grenoble, France	http://www.esrf.fr/
HASYLAB DESY, Deutsches Elektronen-Synchrotron	Hamburg, Germany	http://www.desy.de/
Kurchatov Center for Synchrotron Radiation and Nanotechnology	Moscow, Russian Federation	http://www.kcsr.kiae.ru/en/
LNLS, National Synchrotron Light Laboratory	Campinas, Brazil	http://www.lnls.br/
MAXLab (see also MAX IV project)	Lund, Sweden	http://www.maxlab.lu.se/, http://www.maxlab.lu.se/maxlab/max4/index.html
NSLS, National Synchrotron Light Source (see also NSLS II; under construction)	Brookhaven National Lab., New York, USA	http://www.nsls.bnl.gov/, http://www.bnl.gov/nsls2/
The Photon Factory, KEK	Tsukuba, Japan	http://pfwww.kek.jp/
PLS, Pohang Light Source	Pohang, Korea	http://pal.postech.ac.kr/
SESAME (Synchrotron-light for Experimental Science and Applications in the Middle East)	Allan, Jordan	http://www.sesame.org.jo/
Shanghai Synchrotron Radiation Facility	Shanghai, China	http://ssrf.sinap.ac.cn/english/1/Introduction.htm
SLS, Swiss Light Source	Paul Scherrer Institut, Villigen, Switzerland	http://sls.web.psi.ch/view.php/about/index.html
Soleil	Gif-sur-Yvette, Paris, France	http://www.synchrotron-soleil.fr/portal/page/portal/Accueil
SPring-8, Super Photon Ring	Riken Go, Japan	http://www.spring8.or.jp/
SRRC, Synchrotron Radiation Research Center	Hsinchu City, Taiwan	http://www.nsrrc.org.tw/
SSRL, Stanford Synchrotron Radiation Laboratory	SLAC, California, USA	http://www-ssrl.slac.stanford.edu/
VEPP-3	Novosibirsk, Russia	http://ssrc.inp.nsk.su/

At the sample position, the intensity of the beam, usually focused, is a useful parameter:

$$\text{Intensity} = \text{photons per s per focal spot area.} \qquad (8.1.4.2)$$

Moreover, the horizontal and vertical convergence angles are ideally kept smaller than the mosaic spread, *e.g.* ~1 mrad, so as to measure reflection intensities with optimal peak-to-background ratio.

Producing a focal spot area that is approximately the size of a typical crystal (~0.1 mm) and with a convergence angle ~1 mrad sets a sample acceptance requirement to be met by the X-ray beam and machine emittance. A machine with an emittance that matches the acceptance of the sample greatly assists the simplicity and performance of the beamline optics (mirror and/or monochromator) design. The common beamline optics schemes are shown in Fig. 8.1.4.1.

In addition to the focal spot area and convergence angles, it is necessary to provide the appropriate spectral characteristics. In monochromatic applications, involving the rotating-crystal diffraction geometry, for example, a particular wavelength, λ, and narrow spectral bandwidth, $\delta\lambda/\lambda$, are used. Fig. 8.1.4.2(*a*) shows an example of a monochromatic oscillation diffraction photograph from a rhinovirus crystal as an example recorded at CHESS, Cornell. Fig. 8.1.4.2(*b*) shows the prediction of a white-beam broad-band Laue diffraction pattern from a protein crystal that was recorded at the SRS wiggler, Daresbury, colour-coded for multiplicity.

Table 8.1.4.1 lists the internet addresses of the SR facilities worldwide that currently have macromolecular beamlines. A considerable suite of information on SR and free electron laser (FEL) sources can also be found at http://www.lightsources.org/cms/. Comprehensive statistics for the macromolecular crystal structures from all the various beamlines over the years can be obtained at http://biosync.rcsb.org/.

8.1.5. Evolution of SR machines and experiments

8.1.5.1. First-generation SR machines

The so-called first generation of SR machines were those which were parasitic on high-energy physics operations, such as DESY in Hamburg, SPEAR in Stanford, NINA in Daresbury and VEPP in Novosibirsk. These machines had high fluxes into the X-ray range and enabled pioneering experiments. Parratt (1959) discussed the use of the CESR (Cornell Electron Storage Ring) for X-ray diffraction and spectroscopy in a very perceptive paper. Cauchois *et al.* (1963) conducted *L*-edge absorption spectroscopy at Frascati and were the first to diffract SR with a crystal (quartz). The opening experimental work in the area of biological diffraction was by Rosenbaum *et al.* (1971). In protein crystallography, multiple-wavelength anomalous-dispersion effects (Fig.

8.1.5.1) were used from the onset (Phillips *et al.*, 1976, 1977; Phillips & Hodgson, 1980; Webb *et al.*, 1977; Harmsen *et al.*, 1976; Helliwell, 1977, 1979), and a reduction in radiation damage was seen for high-resolution data collection (Wilson *et al.*, 1983). Historical insights into the performances of those machines, from the current-day perspective, are described in detail, for example, by Huxley & Holmes (1997) at DESY, Munro (1997) at Daresbury, and Doniach *et al.* (1997) at Stanford. A principal limitation was the problem of source movements, which degraded the focusing of the source onto a small crystal or single fibre and thus degraded the intrinsic spectral brightness of the beam; see, for example, Haslegrove *et al.* (1977), who advocated machine shifts dedicated to SR as a working compromise with the high-energy physicists. Some possible applications that were discussed were unfulfilled until brighter sources became available. The two-wavelength crystallography phasing method of Okaya & Pepinsky (1956) (see also Hoppe & Jakubowski, 1975) and the three-wavelength method of Herzenberg & Lau (1967), as well as the implementation of the algebraic method of Karle (1967, 1980, 1989, 1994), awaited more stable beams, which had to be rapidly and easily tunable over a fine bandpass (ideally 10^{-4}). Experiments to define the anomalous-dispersion coefficients, including dichroism effects, at a large number of wavelengths at several example absorption edges in a variety of crystal structures were conducted at SPEAR (Phillips *et al.*, 1978; Templeton *et al.*, 1980, 1982; Templeton & Templeton, 1985). Large values of f'' were identified at 'white lines', *i.e.* regions of the elemental absorption with pronounced effects (*e.g.* see Fig 8.1.5.1*a*). Values of f' over a continuum of wavelengths in a real compound (*i.e.*, not a metal in the gas phase) (Fig. 8.1.5.1*b*) were explored in a profile approach (now called DAFS, diffraction anomalous fine structure) by Arndt *et al.* (1982) at the newly commissioned SRS, the first dedicated second-generation SR source (see Section 8.1.5.2).

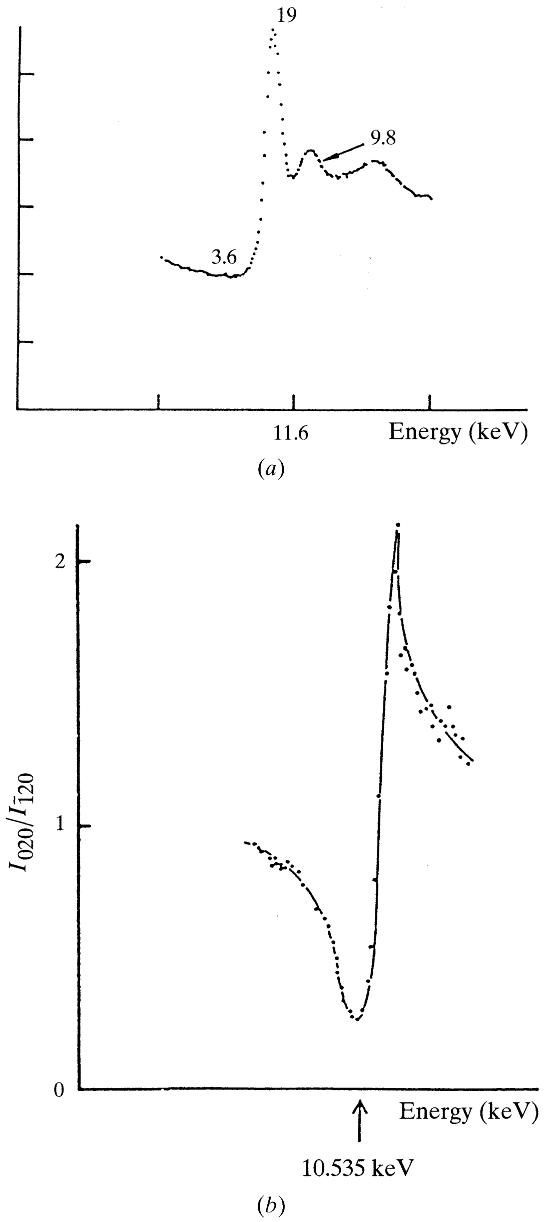


Figure 8.1.5.1
Anomalous dispersion. (*a*) f'' as represented by an absorption spectrum [Pt L_{III} edge for $K_2Pt(CN)_4$ as the example with 19 electrons for f'' at the peak of that 'white line', with pre-edge and post-edge f'' electron values also indicated] (Helliwell, 1984). Reproduced with the permission of the Institute of Physics. (*b*) f' as estimated by a continuous polychromatic profile method. Reproduced with permission from *Nature* (Arndt *et al.*, 1982). Copyright (1982) MacMillan Magazines Limited.

8.1.5.2. *Second-generation dedicated machines*

The building of dedicated X-ray sources began with the SRS at Daresbury, which came online in 1980, having followed the NINA synchrotron (closed in 1976) and the associated Synchrotron Radiation Facility at Daresbury. Elsewhere in the world, LURE (Lemonnier *et al.*, 1978) and CHESS at Cornell were building up their SR macromolecular crystallography operations in the late 1970s and early 1980s, and the NSLS in Brookhaven and the Photon Factory (PF) in Japan were both under construction. The NSLS and the PF came online in 1983 and 1984, respectively. Thus, there was a rapid increase in the number of operating machines and beamlines worldwide in the X-ray region for protein crystallography. There were teething problems with the SRS with the radio-frequency cavity window problem, interrupting operation for many months in 1983, and at the NSLS in its early period due to vacuum-chamber problems. Pioneering experiments continued and blossomed. Seminal work ensued in virus crystallography [Rossmann & Erickson (1983) at Hamburg and Daresbury; and Usha *et al.* (1984) at LURE], Laue diffraction for time-resolved protein crystallography [Moffat *et al.* (1984) at CHESS; Helliwell (1984, 1985) at the SRS; Cruickshank *et al.* (1987, 1991); Hajdu, Machin *et al.* (1987); Helliwell *et al.* (1989); Bourenkov *et al.* (1996); Neutze & Hajdu (1997)], enzyme catalysis in the crystal [Hajdu, Acharya *et al.* (1987) at the SRS], MAD [Phillips *et al.* (1977); Einspahr *et al.* (1985); Hendrickson (1985); Hendrickson *et al.* (1989) at SPEAR, the SRS and the PF; Guss *et al.* (1988) at SPEAR; Kahn *et al.* (1985) at LURE; Korszun (1987) at CHESS; Mukherjee *et al.* (1989) and Peterson *et al.* (1996) at the SRS; Hädener *et al.* (1999) at the SRS and the ESRF, to cite a few experiments], protein crystallography involving isomorphous replacement with optimized anomalous scattering [Baker *et al.* (1990) at the SRS; Dumas *et al.* (1995) at LURE], small crystals [Hedman *et al.* (1985) at the SRS] and diffuse scattering with SR [Doucet & Benoit (1987); Caspar *et al.* (1988); Glover *et al.* (1991)]. Table 8.1.5.1 shows the impact of the SRS in protein (*i.e.* macromolecular) crystallography integrated over its whole lifetime.

8.1.5.3. *Third-generation high spectral brightness machines*

As early as 1979, there were discussions on planning a proposal for a high spectral brightness, insertion-device-driven European synchrotron-radiation (ESR) source. A wide variety of discussion

Table 8.1.5.1
Structures in the Protein Data Bank (PDB) for which data were collected at the SRS

The data presented here were compiled in December 2009 (see http://biosync.rcsb.org/biosync_regions/SyncEurope.html#SRS) and are likely to be reasonably complete since the SRS closed operations in August 2008. The SRS has delivered 3.6% of the total of 38 650 macromolecular crystal structures determined using radiation from synchrotrons around the world as of December 2009. The ESRF third-generation source, in comparison, integrated over about half as many years, but about two to three times more beamlines, has delivered 15.6% of the structures, *i.e.* at a rate therefore about three to four times greater than the second-generation SRS.

	Station						
Year	10.1	14.1	14.2	7.2	9.5	9.6	Not known
1995	0	0	0	5	6	19	3
1996	0	0	0	6	11	33	0
1997	0	0	0	11	29	43	0
1998	0	0	0	26	32	35	0
1999	0	0	0	28	13	45	4
2000	0	1	0	28	17	60	3
2001	0	13	9	9	16	47	2
2002	0	13	21	7	7	59	2
2003	0	27	38	3	8	41	3
2004	3	40	42	5	2	36	2
2005	18	34	36	1	4	47	1
2006	22	32	22	1	1	23	0
2007	21	37	21	0	0	11	1
2008	51	15	15	2	0	20	1
2009	14	12	5	1	0	3	0
Total	129	224	209	133	146	522	22

documents and workshops, and the ESR Project (ESRP) led by B. Buras and based in Geneva at CERN, culminated in the so-called 'Red Book' in 1987, the *ESRF Foundation Phase Report* (1987), totalling some 1000 pages of machine, beamline and experimental specifications and costs. This, then, was the progenitor of the third-generation sources, characterized by their high energy and high spectral brightness, tailored to optimized undulator emission in the 1 Å range. Actually, the ESRF machine energy was initially set at 5 GeV, but increased to 6 GeV to optimize the production of 14.4 keV photons to better match the nuclear scattering experiments proposed initially by Mossbauer in 1975. Proposals for the US machine, the Advanced Photon Source at 7 GeV, and the Japanese 8 GeV SPring-8 machine followed, with the higher machine energy enhancing the X-ray tuning range of undulators. Thus, MAD tuning-based techniques were facilitated with these machines and studies involving yet-smaller samples (crystals, single fibres or tiny liquid aliquots) or very large unit cells were enabled. As a result, micron-sized protein crystals as well as huge multi-macromolecular biological structures (of large viruses, for example) also became routinely accessible.

8.1.5.4. New national SR machines

Today a variety of enhanced national SR machines have been built. In Switzerland there is the SLS, in the UK there is DIAMOND and in France there is SOLEIL. These machines are more tailored to the bulk of a country's user needs, distinct from the special provisions at the ESRF. The different countries' SR needs, of course, have many aspects in common, with some historical biases. The new sources are, in essence, characterized by high spectral brightness, *i.e.*, low emittance. The 2 GeV SR source ELETTRA in Trieste, the MAXII machine in LUND and the Brazilian Light Source are already operational. In many ways, national sources like the SRS, LURE, DORIS and so on fuelled the case and specification for the ESRF. Now the developments at the ESRF, including high harmonic emission of undulators *via* magnet shimming (Elleaume, 1989) and narrow-gap undulator operation (Elleaume, 1998), are fuelling ideas and the specification of what is possible in the new national SR sources.

8.1.5.5. X-ray free-electron lasers (XFELs)

In terms of the evolution of X-ray sources, X-ray FELs are being constructed at DESY in Hamburg (Brinkmann *et al.*, 1997), at SLAC (Winick, 1995) and at Spring8. Compared to SR, one will have a transversely fully coherent beam, a larger average spectral brightness and, in particular, pulse lengths of ~10 fs full width at half-maximum with eight to ten orders of magnitude larger peak spectral brightness. Such a machine is based on a linear accelerator (linac)-driven XFEL utilizing a linear collider installation (*e.g.*, for a high-energy physics centre-of-mass energy capability of 500 GeV). For this machine there is a 'switchyard' distributing the electrons in a beam to different undulators from which the X-rays are generated in the range 0.1 to ~12 keV. The anticipated r.m.s. opening angle would be 1 mrad and the source diameter would be 20 μm. This source of X-rays would then compete in time resolution with laser-pulse-generated X-ray beams [see Helliwell & Rentzepis (1997) for a survey of that work and a comparison with synchrotron radiation] and would also have higher pulse flux. Coherent methods in the X-ray sciences have been extensively reviewed by Nugent (2009).

8.1.6. SR instrumentation

The divergent continuum of X-rays from the source must be intercepted by the sample cross-sectional area. The crystal sample acceptance, as seen above, is a good way to illustrate to the machine designer the sort of machine emittances required. Likewise, the beamline optics, mirrors and monochromators should not degrade the X-ray beam quality. Mirror surface and shape finish have improved a great deal in the last few decades; slope errors of mirrors, even for difficult shapes like polished cylinders, which on bending give a toroidal reflecting surface, are now around 1 arc second (5.5 μrad) for a length of 1 m. Thus, over focusing distances of 10–20 m, say, the focal-spot smearing contribution from this is 55–110 μm, important for focusing onto small crystals. Further optics developments (*e.g.* Fresnel optics) have yielded micron focus beams and smaller, and are being applied to studying ever-smaller crystals in macromolecular crystallography and obviously have a variety of other diffraction and spectroscopy applications [for a review, see Riekel (2000)]. The choice of materials has evolved, too, from the relatively easy-to-work-with and -finish fused quartz to silicon; silicon having the advantageous property that at liquid-nitrogen temperature the expansion coefficient is zero (Bilderback, 1986). This has been of particular advantage in the cooling of silicon monochromators at the ESRF, where the heat loading on optics is very high. An alternative approach with the rather small X-ray beams from undulators is the use of transparent monochromator crystals made of diamond, which is a robust material with the additional advantage of transparency, thus allowing multiplexing of stations, one downstream from the other, fed by one straight section of one or more undulator designs. For a review of the ESRF beamline optics, see Freund (1996); for reviews of the macromolecular crystallography programmes at the ESRF, see Miller (1994), Branden (1994) and Lindley (1999), as well as the *ESRF Foundation Phase Report* (1987). See also Helliwell (1992), Chapter 5.

Detectors have improved enormously. The early days of SR use saw considerable reliance on photographic film, as well as single-counter four-circle diffractometers. Evolution of area detectors, in particular, has been considerable and impressive, and in a variety of technologies. Gas detectors, *i.e.*, the multiwire proportional chamber (MWPC), were invented and developed through various generations and types [Charpak (1970); for reviews of their use at SR sources, see *e.g.* Lewis (1994) and Fourme (1997)]. MWPCs have the best detector quantum efficiency (DQE) of the area detectors, but there are limitations on count rate (local and global) and their use at wavelengths less than ~1 Å is restricted due to geometric image parallax effects. The most popular devices at present are charge coupled devices (CCDs) [see Tate *et al.* (1995), Allinson (1994), Gruner & Ealick (1995) and Westbrook & Naday (1997) for details of their development]. Image plates (IPs) were popular during the late 1980s and early to mid-1990s, mainly, but not exclusively, with online scanners, notably the MAR Research devices. IPs are also used in a Weissenberg geometry [see Sakabe (1983, 1991) and Sakabe *et al.* (1995), and for a recent review see Amemiya (1997)]. IPs and CCDs are complementary in performance, especially with respect to size and duty cycle; IPs are larger, *i.e.*, with many resolution elements possible, but are slower to read out than CCDs. Both are capable of imaging well at wavelengths shorter than 1 Å and with high count rates. Both have overcome the tedium of chemical development of film. Other detectors needed for crystallography include those for monitoring the beam intensity; these must not interfere with the beam collimation, and yet must monitor the beam downstream of the collimator (Bartunik *et al.*, 1981); also needed are fluorescence detectors for setting the wavelength for optimized anomalous-scattering applications (see Cianci *et al.*, 2005).

Most recently, an area-detector development has been the so-called pixel detector. This is made of silicon cells, each 'bump bonded' onto associated individual electronic readout chains. Thus, extremely high count rates are possible. These devices can then combine the attributes of large image plate sensitive areas with the fast readout of CCDs, along with high count-rate capability and so on. Devices and prototypes have been developed at Princeton/Cornell (Eikenberry *et al.*, 1998), Berkeley/San Diego (Beuville *et al.* 1997) and Imperial College, London (Hall, 1995), and are now in use at the SLS (Broennimann *et al.*, 2006).

Provision of robotics for sample mounting on the synchrotron beamlines has been increasingly deployed in the last decade, improving efficiency and ease of use, often coupled with remote access (*e.g.* see Gonzalez *et al.*, 2005) and telepresence (*e.g.* see Warren *et al.*, 2008).

8.1.7. SR monochromatic and Laue diffraction geometry

In the utilization of SR, both Laue and monochromatic modes are important for data collection. The unique geometric and spectral properties of SR render the treatment of diffraction geometry different from that for a conventional X-ray source.

8.1.7.1. Laue geometry: sources, optics, sample reflection bandwidth and spot size

Laue geometry involves the use of the polychromatic SR spectrum as transmitted through the beryllium window that is used to separate the apparatus from the machine vacuum. There is useful intensity down to a wavelength minimum of $\sim\lambda_c/5$, where λ_c is the critical wavelength of the magnet source in the case of bending magnets (BMs) and wavelength shifters. The maximum wavelength is typically ≥ 3 Å; however, if the crystal is mounted in a capillary, then the glass absorbs the wavelengths beyond ~2.5 Å (Helliwell, 2004).

The bandwidth on BMs and wigglers can be limited by a reflecting mirror at grazing incidence, whereby the minimum wavelength in the beam can be sharply defined to aid the accurate definition of the Laue spot multiplicity. The mirror can also be used to focus the beam at the sample. The maximum-wavelength limit can be truncated by use of aluminium absorbers of varying thickness or a transmission mirror (Lairson & Bilderback, 1982; Cassetta *et al.*, 1993).

The measured intensity of individual Laue diffraction spots depends on the wavelength at which they are stimulated. The problem of wavelength normalization is treated by a variety of methods. These include:

(i) use of a monochromatic reference data set;
(ii) use of symmetry equivalents and multiple measurements in the Laue data set recorded at different wavelengths;
(iii) calibration with a standard sample, such as a silicon crystal.

Each of these methods produces a 'λ curve' describing the relative strength of spots measured at various wavelengths. Most Laue diffraction data are now recorded on CCDs or IPs. The greater sensitivity of these detectors (expressed as the DQE), especially for weak signals, has greatly increased the number of Laue exposures recordable per crystal (*e.g.* Nieh *et al.*, 1999). Thus, multiplet deconvolution procedures, based on the recording of reflections stimulated at different wavelengths and with different relative intensities, have become possible (Campbell & Hao, 1993; Ren & Moffat, 1995*b*; Nieh *et al.*, 1999). Data quality and completeness have improved considerably.

Narrow-bandpass beams, *e.g.* $\delta\lambda/\lambda \leq 0.2$, are used for enhancing the signal-to-noise ratio. Such bandwidths are produced generally by an X-ray undulator, whereby *e.g.* 10–100 periods should ideally yield a bandwidth behind a pinhole of $\delta\lambda/\lambda \simeq 0.1$–$0.01$. In these cases, wavelength normalization is more difficult, because the actual spectrum over which a reflection is integrated is rapidly varying in intensity; nevertheless, high-order Chebychev polynomials are successful (Ren & Moffat, 1995*a*; Artz *et al.*, 1999).

The spot bandwidth is determined by the mosaic spread and horizontal beam divergence (since $\gamma_H > \gamma_V$) as

$$(\delta\lambda/\lambda) = (\eta + \gamma_H)\cot\theta, \qquad (8.1.7.1)$$

where η is the sample mosaic spread, assumed to be isotropic, and

γ_H is the horizontal cross-fire angle, which in the absence of focusing is $(x_H + \sigma_H)/P$, where x_H is the horizontal sample size, σ_H is the horizontal source size and P is the sample to the tangent-point distance. This is similar for γ_V in the vertical direction. Generally, at SR sources, σ_H is greater than σ_V. When a focusing-mirror element is used, γ_H and/or γ_V are convergence angles determined by the focusing distances and the mirror aperture.

The size and shape of the diffraction spots vary across the detector image plane. The radial spot length is given by convolution of Gaussians as

$$\left(L_R^2 + L_c^2 \sec^2 2\theta + L_{\rm PSF}^2\right)^{1/2} \qquad (8.1.7.2)$$

and tangentially by

$$\left(L_T^2 + L_c^2 + L_{\rm PSF}^2\right)^{1/2}, \qquad (8.1.7.3)$$

where L_c is the size of the X-ray beam (assumed to be circular) at the sample, $L_{\rm PSF}$ is the detector point-spread factor,

$$L_R = D \sin(2\eta + \gamma_R) \sec^2 2\theta, \qquad (8.1.7.4)$$

$$L_T = D(2\eta + \gamma_T) \sin\theta \sec 2\theta, \qquad (8.1.7.5)$$

and

$$\gamma_R = \gamma_V \cos\psi + \gamma_H \sin\psi, \qquad (8.1.7.6)$$

$$\gamma_T = \gamma_V \sin\psi + \gamma_H \cos\psi, \qquad (8.1.7.7)$$

where ψ is the angle between the vertical direction and the radius vector to the spot (see Andrews *et al.*, 1987). For a crystal that is not too mosaic, the spot size is dominated by L_c and $L_{\rm PSF}$. For a mosaic or radiation-damaged crystal, the main effect is a radial streaking arising from η, the sample mosaic spread.

8.1.7.2. Monochromatic SR beams: optical configurations and sample rocking width

A wide variety of perfect-crystal monochromator configurations are possible and have been reviewed by various authors (Hart, 1971; Bonse *et al.*, 1976; Hastings, 1977; Kohra *et al.*, 1978). Since the reflectivity of perfect silicon and germanium is effectively 100%, multiple-reflection monochromators are feasible and permit the tailoring of the shape of the monochromator resolution function, harmonic rejection and manipulation of the polarization state of the beam. Two basic designs are in common use. These are the bent single-crystal monochromator of triangular shape (Fig. 8.1.4.1*a*) and the double-crystal monochromator (Fig. 8.1.4.1*b*).

8.1.7.2.1. Curved single-crystal monochromator

In the case of the single-crystal monochromator, the actual curvature employed is very important in the diffraction geometry. For a point source and a flat monochromator crystal, there is a gradual change in the photon wavelength selected from the white beam as the length of the monochromator is traversed (Fig. 8.1.7.1*a*). For a point source and a curved monochromator crystal, one specific curvature can compensate for this variation in incidence angle (Fig. 8.1.7.1*b*). The reflected spectral bandwidth is then at a minimum; this setting is known as the 'Guinier position'. If the curvature of the monochromator crystal is increased further, a range of photon wavelengths, $(\delta\lambda/\lambda)_{\rm corr}$, is selected along its length so that the rays converging towards the focus have a correlation of photon wavelength and direction (Fig. 8.1.7.1*c*). The effect of a finite source is to cause a change in incidence angle at the monochromator crystal, so that at the focus there is a photon-wavelength gradient across the width of the focus (for all curvatures) (Fig. 8.1.7.1*d*). The use of a slit in the focal plane is akin to placing a slit at the tangent point to limit the source size.

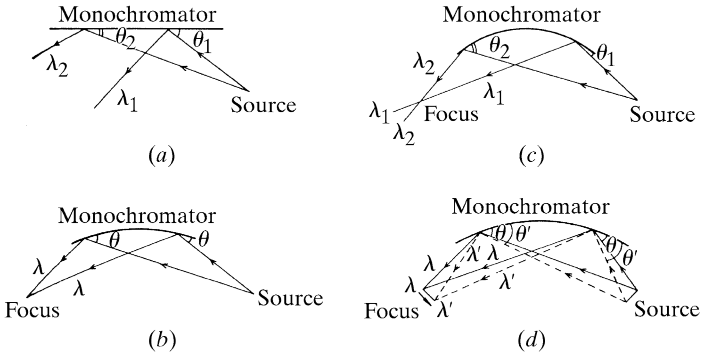


Figure 8.1.7.1
Single-crystal monochromator illuminated by SR. (*a*) Flat crystal. (*b*) Guinier setting. (*c*) Overbent crystal. (*d*) Effect of source size (shown at the Guinier setting for clarity). From Helliwell (1984). Reproduced with the permission of the Institute of Physics.

8.1.7.2.2. Double-crystal monochromator

The double-crystal monochromator with two parallel or nearly parallel perfect crystals of germanium or silicon is a common configuration. The advantage of this is that the outgoing monochromatic beam is parallel to the incoming beam, although it is slightly displaced vertically by an amount $2d\cos\theta$, where d is the perpendicular distance between the crystals and θ is the monochromator Bragg angle (unless the second crystal is unconnected to the first, in which case it can be translated as well to compensate for that). The monochromator can be rapidly tuned, since the diffractometer or camera need not be re-aligned significantly in a scan across an absorption edge. Since the rocking width of the fundamental is broader than the harmonic reflections, the strict parallelism of the pair of the crystal planes can be 'detuned', so that the harmonic can be rejected with little loss of the fundamental intensity. The spectral spread in the reflected monochromatic beam is determined by the source divergence accepted by the monochromator, the angular size of the source and the monochromator rocking width (see Fig. 8.1.7.2). The double-crystal monochromator is often used with a toroidal focusing mirror; the functions of monochromatization are then separated from the focusing (Hastings *et al.*, 1978).

8.1.7.2.3. Widening of monochromatic wavelength range provision

Users and facilities have been driven to use an ever-widening practical range of X-ray wavelengths. Longer wavelengths in protein crystallography have been explored up to 2.6 Å (Helliwell, 1993, 1997*a*, 2004; Polikarpov *et al.*, 1997; Teplyakov *et al.*, 1998; Cianci *et al.*, 2001, 2008), and even softer wavelengths have been under active development to utilize the uranium M edge at 3.326 and 3.490 Å (Liu *et al.*, 2001), and the S K edge for anomalous dispersion (Stuhrmann & Lehmann, 1994). At very short (~0.5 Å) and ultra-short (~0.3 Å) wavelengths, a high machine energy yields copious flux output; pilot studies have been conducted in protein crystallography at CHESS (Helliwell *et al.*, 1993), at the ESRF (Schiltz *et al.*, 1997) and at the NSLS (Jakoncic *et al.*, 2006). Notably Fourme and colleagues have exploited the condensed angular distribution of the diffraction pattern at such wavelengths to pass the beam through the narrow

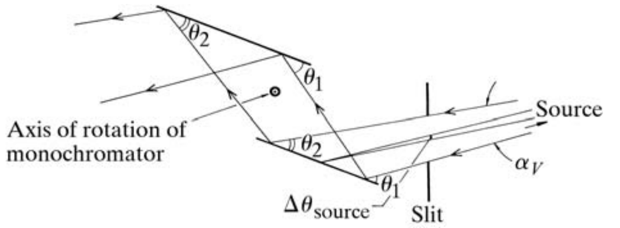


Figure 8.1.7.2
Double-crystal monochromator illuminated by SR. The contributions of the source divergence, α_V [less than or equal to γ^{-1}, equation (8.1.2.4), depending on the monochromator vertical entrance slit aperture; see also Colapietro *et al.* (1992)], and angular source size, $\Delta\theta_{\rm source}$, to the range of energies reflected by the monochromator are shown. From Helliwell (1984). Reproduced with the permission of the Institute of Physics.

exit window of a high-pressure cell for protein-crystal high-pressure biophysical studies (Fourme *et al.*, 2001; Girard *et al.*, 2007). Such developments interact closely with machine and beamline specifications.

8.1.7.2.4. Crystal sample rocking width

The rocking width of a reflection depends on the horizontal and vertical beam divergence or convergence (after due account for collimation is taken), γ_H and γ_V, the spectral spreads $(\delta\lambda/\lambda)_{\rm conv}$ and $(\delta\lambda/\lambda)_{\rm corr}$, and the mosaic spread, η. We assume that the mosaic spread η is $\gg \omega$, the angular broadening of a reciprocal-lattice point (relp) due to a finite sample. In the case of synchrotron radiation, γ_H and γ_V are usually widely asymmetric. On a conventional source, usually $\gamma_H \simeq \gamma_V$. Two types of spectral spread occur with synchrotron (and neutron) sources. The term $(\delta\lambda/\lambda)_{\rm conv}$ is the spread that is passed down each incident ray in a divergent or convergent incident beam; the subscript refers to the conventional source type. This is because it is similar to the $K\alpha_1, K\alpha_2$ line widths and separation. At the synchrotron, this component also exists and arises from the monochromator rocking width and finite-source-size effects. The term $(\delta\lambda/\lambda)_{\rm corr}$ is special to the synchrotron or neutron case. The subscript 'corr' refers to the fact that the ray direction can be correlated with the photon or neutron wavelength. In this most general case, and for one example of a $(\delta\lambda/\lambda)_{\rm corr}$ arising from the horizontal ray direction correlation with photon energy and the case of a horizontal rotation axis, the rocking width φ_R of an individual reflection is given by

$$\varphi_R = \left\{L^2\left[(\delta\lambda/\lambda)_{\rm corr}d^{*2} + \zeta\gamma_H\right]^2 + \gamma_V^2\right\}^{1/2} + 2\varepsilon_s L, \qquad (8.1.7.8)$$

where

$$\varepsilon_s = (d^*\cos\theta/2)\left[\eta + (\delta\lambda/\lambda)_{\rm conv}\tan\theta\right] \qquad (8.1.7.9)$$

and L is the Lorentz factor, $1/\left(\sin^2 2\theta - \zeta^2\right)^{1/2}$.

The Guinier setting of an instrument (curved crystal monochromator case, Fig. 8.1.7.1*b*) gives $(\delta\lambda/\lambda)_{\rm corr} = 0$. The equation for φ_R then reduces to

$$\varphi_R = L\left[\left(\zeta^2\gamma_H^2 + \gamma_V^2/L^2\right)^{1/2} 2\varepsilon_s\right] \qquad (8.1.7.10)$$

(from Greenhough & Helliwell, 1982). For example, for $\zeta = 0$, $\gamma_V = 0.2$ mrad (0.01°), $\theta = 15°$, $(\delta\lambda/\lambda)_{\rm conv} = 1 \times 10^{-3}$ and $\eta = 0.8$ mrad (0.05°), then $\varphi_R = 0.08°$. But φ_R increases as ζ increases [see Greenhough & Helliwell (1982), Table 5]. In the rotation/oscillation method as applied to protein and virus crystals, a small angular range is used per exposure. For example, the maximum rotation range per image, $\Delta\varphi_{\max}$, may be 1.5° for a protein and 0.4° or so for a virus. Many reflections will be only partially stimulated over the exposure. It is important, especially in the virus case, to predict the degree of penetration of the relp through the Ewald sphere. This is done by analysing the interaction of a spherical volume for a given relp with the Ewald sphere. The radius of this volume is given by

$$E \simeq \varphi_R/2L \qquad (8.1.7.11)$$

(Greenhough & Helliwell, 1982).

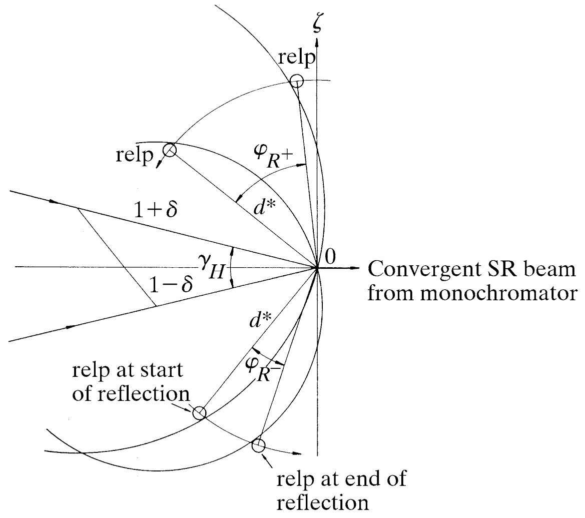

ζ
relp
relp
φR+
d*
1+δ
γH
1−δ
0
Convergent SR beam from monochromator
d*
φR−
relp at start of reflection
relp at end of reflection

Figure 8.1.7.3
The rocking width of an individual reflection for the case of Fig. 8.1.7.1(*c*) and a vertical rotation axis. From Greenhough & Helliwell (1982).

In Fig. 8.1.7.3, the relevant parameters are shown. The diagram shows $(\delta\lambda/\lambda)_{\rm corr} = 2\delta$ in a plane, usually horizontal with a perpendicular (vertical) rotation axis, whereas the formula for φ_R above is for a horizontal axis. This is purely for didactic reasons since the interrelationship of the components is then much clearer.

The limits of protein-crystal rocking widths have been explored by Helliwell (1988), Colapietro *et al.* (1992) and Snell (1995), whereby arc-second crystal precision has been observed at room temperature, *i.e.* unfrozen protein crystals. Special analysis software (Lovelace *et al.*, 2000) has been written to extract such precise crystal mosaicity values and the experiment obviously requires stringent (usually undulator) collimation. Exact comparisons of mosaicity values need reflection indexing, *e.g.* see Snell *et al.* (1995).

8.1.8. Scientific utilization of SR in protein crystallography

There are a myriad of applications and results of the use of SR in crystallography. Helliwell (1992) gave an extensive survey and tabulations of SR and macromolecular crystallography applications; Chapter 9 therein concentrates on anomalous scattering and Chapter 10 on high resolution, large unit cells, small crystals, weak scattering efficiency and time-resolved data collection. The field has expanded so dramatically, in fact, that an equivalent survey today would be vast. Table 8.1.4.1 lists the web pages of the facilities, where the specifications and details of the beamlines can be found (*e.g.* all the publications at Daresbury in the protein crystallography area organized by beamline instrument are to be found at http://dlwebres.dl.ac.uk/dl_public/publications/index.jsp). The examples below cite extreme cases of the large unit cell (virus and multi-macromolecular) cases, weak anomalous-scattering signal in MAD, fast time-resolved Laue

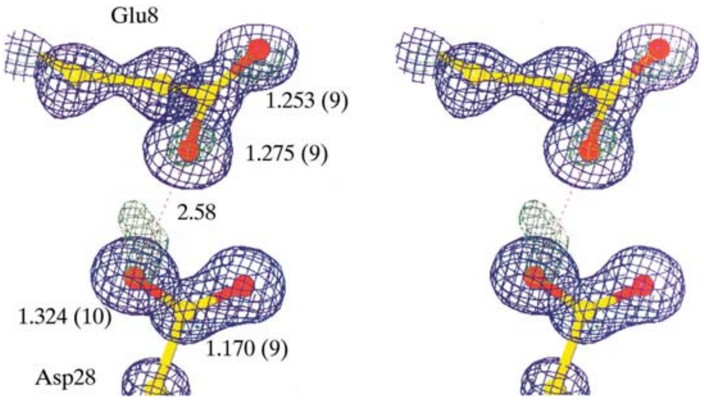


Figure 8.1.8.1
Determination of the protonation states of carboxylic acid side chains in proteins *via* hydrogen atoms and resolved single and double bond lengths. After Deacon *et al.* (1997) using CHESS. Reproduced by permission of The Royal Society of Chemistry.

studies and the ultra-high-resolution and even valence-density structure determinations to date. Another phasing technique involving multiple (‘*n*-beam’) diffraction is also being applied to proteins [Weckert & Hümmer (1997) at the ESRF and the NSLS]. These examples at least indicate the present bounds of capability of the various sub-fields of SR and macromolecular crystallography.

8.1.8.1. Atomic and ultra-high-resolution macromolecular crystallography

The use of high SR intensity, cryo-freezing of a protein crystal to largely overcome radiation damage and sensitive, automatic area detectors (CCDs and/or image plates) is allowing diffraction data to be recorded at resolutions equivalent to smaller-molecule (chemical) crystallography. In a growing number of protein crystal structure studies, atomic resolution (1.2 Å or better) is achievable (Dauter *et al.*, 1997). The ‘X-ray data to parameter’ ratio can be favourable enough for single and double bonds, *e.g.* in carboxyl side chains, to be resolved [Fig. 8.1.8.1; Deacon *et al.* (1997) for concanavalin A at 0.94 Å resolution]. Along with this bond-distance precision, one can see the reactive proton directly. This approach complements H/D exchange neutron diffraction studies. Neutron studies have expanded in scope by employing Laue geometry in a synergistic development with SR Laue diffraction (Helliwell & Wilkinson, 1994; Helliwell, 1997*b*; Habash *et al.*, 1997, 2000); a comprehensive survey of neutron macromolecular crystallography, instruments and results in determining the atomic details of protonation and hydration has been given by Blakeley (2009). In particularly well ordered protein structure cases, valence-electron-density descriptions are possible for those atoms with *B* factors ~<3 Å^2 [see *e.g.* Guillot *et al.* (2008) and Luger (2007)]. The scope and accuracy of protein crystal structures has been transformed. A diffraction-component precision index for characterizing the overall precision of protein structures has been given by Cruickshank (1999) and cast in terms of experimental parameters by Blow (2002).

8.1.8.2. Small crystals

Compensating for small crystal sample volume by increasing the intensity at the sample has been of major interest from the outset, and tests showed that the use of ~20 µm-sized samples is feasible (Hedman *et al.*, 1985). Third-generation high spectral brightness sources were optimized for this application *via* micron-sized focal spot beams, as described in the *ESRF Foundation Phase Report* (1987). Applications of the ESRF microfocus beamline diffractometer (Perrakis *et al.*, 1999) include as an example the determination of the structure of the bacteriorhodopsin crystal at high resolution from micro-crystals (Pebay-Peyroula *et al.*, 1997). Experiments using extremely thin plates involving only 1000 protein molecular layers are described by Mayans & Wilmanns (1999) on the BW7B wiggler beamline at DESY, Hamburg. A variety of small crystals and SR, including tabulated sample scattering efficiencies, can be found in Helliwell (1992), pp. 410–414 and more recently in Riekel *et al.* (2005). The ESRF Upgrade (Fig 8.1.2.2) will push this into the sub-micron territory.

8.1.8.3. Time-resolved macromolecular crystallography

Time-resolved SR Laue diffraction of light-sensitive proteins, such as CO Mb studied with sub-nanosecond time resolution in pump–probe experiments (see Srajer *et al.*, 1996), have shown direct structural changes as a function of time. Enzymes, likewise, are being studied directly by time-resolved methods *via* a variety of reaction initiation methods, including pH jump, substrate diffusion and light flash of caged compounds pre-equilibrated in the crystal. Flash freezing is increasingly used to trap molecular structures at optimal times in a reaction determined either by microspectrophotometry or repeated Laue ‘flash photography’. Enzyme reaction rates can be altered through site-directed mutagenesis (*e.g.* see Niemann *et al.*, 1994; Helliwell *et al.*, 1998) and matched to diffraction-data acquisition times. For overviews, see the books edited by Cruickshank *et al.* (1992) and Helliwell & Rentzepis (1997), the recent review by Bourgeois & Weik (2009) and the companion chapter by Moffat (Chapter 8.2 in this volume).

8.1.8.4. Multi-macromolecular complexes

There is now a wealth of results in this field and the reader must be referred to books such as Liljas *et al.* (2009) for a detailed exposition. Multi-macromolecular complexes, such as viruses (Rossmann *et al.*, 1985; Acharya *et al.*, 1989; Liddington *et al.*, 1991) (Fig. 8.1.8.2), the nucleosome (Luger *et al.*, 1997), light-harvesting complex (McDermott *et al.*, 1995) and the 13-subunit membrane-bound protein cytochrome *c* oxidase (Tsukihara *et al.*, 1996), and large-scale molecular assemblies like muscle (Holmes, 1998) are very firmly recognizable as biological entities whose crystal structure determinations relied on SR. These single-crystal structure determinations involved extremely large unit cells and became tractable despite very weak scattering strength. The crystals often showed extreme sensitivity to radiation (hundreds, even a thousand, crystals have been used to constitute a single data set). Cryocrystallography radiation protection is now used extensively in crystallographic data collection and was critical for work with ribosome crystals (Hope *et al.*, 1989); SR has been essential for these structure determinations (see *e.g.* Yonath, 1992; Yonath *et al.*, 1998; Ban *et al.*, 1998; Wimberley *et al.*, 2000; Noller, 2005). A very large multi-protein complex solved using data from the Daresbury SRS wiggler is the F$_1$ ATPase structure (Fig. 8.1.8.3), for which a share in the Nobel Prize for Chemistry in 1997 was awarded to John Walker in Cambridge. The structure (Abrahams *et al.*, 1994; Abrahams & Leslie, 1996) and the amino-acid sequence data, along with fluorescence microscopy, show how biochemical energy is harnessed to drive the proton pump across biological membranes, thus corroborating hypotheses about this process made over many years. This

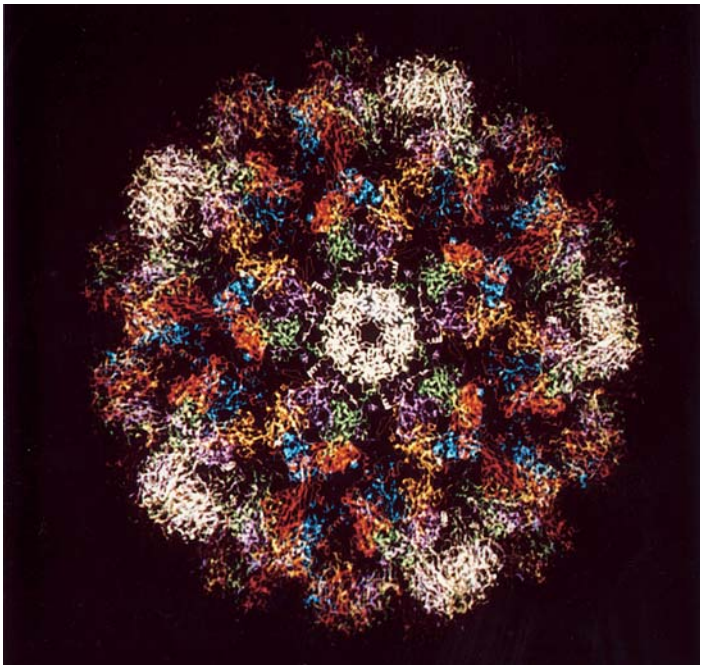


Figure 8.1.8.2
A view of SV40 virus (based on Liddington *et al.*, 1991) determined using data recorded at the SRS wiggler station 9.6 (Fig. 8.1.4.1*a*).

study, made tractable by the SRS wiggler high-intensity protein crystallography station (Fig. 8.1.4.1), illustrates the considerable further scope that became possible with third-generation SR sources, such as the 780 Å diameter blue tongue virus (Grimes *et al.*, 1997, 1998) and the nucleosome core particle (Luger *et al.*, 1997). Spectacular progress has been made in the structural biology of photosynthesis using SR sources, which is not only yielding answers on this vital natural process but also stimulating much research to help address artificial energy sources based on this natural system. This research could have profound climate-change impacts with 'greener energy sources'. One such structure is the Photosystem II (PSII; see Fig. 8.1.8.4). This topic has been reviewed recently by Barber (2009); the atomic details of the Mn$_4$Ca cluster of PSII have proved to be especially challenging to X-ray study (Yano *et al.*, 2005). These large-scale molecular assemblies often combine electron-microscope and diffraction techniques with SR X-ray crystallography and diffraction for low-to-high resolution detail, respectively.

8.1.8.5. Optimized anomalous dispersion (MAD), improved multiple isomorphous replacement (MIR) data and 'structural genomics'

Rapid protein structure determination *via* the MAD method, notably involving seleno protein variants (Hendrickson *et al.*, 1990) as well as xenon pressure derivatives (see *e.g.* Schiltz *et al.*, 1997; Cianci *et al.*, 2001), and improved heavy-atom isomorphous replacement data are removing a major bottleneck in protein crystallography: that of phase determination. Databases of successful heavy-atom compounds (see *e.g.* Sugahara *et al.*, 2009) have been compiled and are increasingly sophisticated. Overall, as the number of protein structures in the Protein Data Bank doubles every few years, the possibility of considering whole genome-level structure determinations arises (Chayen *et al.*, 1996; Chayen & Helliwell, 1998). The human genome comprises some 35 000 genes. Of these, some 40% are coding for membrane-bound proteins, which are more difficult to crystallize. Since a MAD protein crystal structure currently requires less than 1 day of SR BM beamtime, the new 'bottlenecks' are protein production and crystallization. Thus, structural genomics projects have established 'pipelines' for protein structure determination with a view to creating a complete 'protein folds space'. This approach, along with homology modelling and genetic alignment techniques, opens the immense potential for structural genomics to yield huge numbers of experimentally derived protein structures and thereby a much better basis for understanding and controlling disease through structure-based drug design and

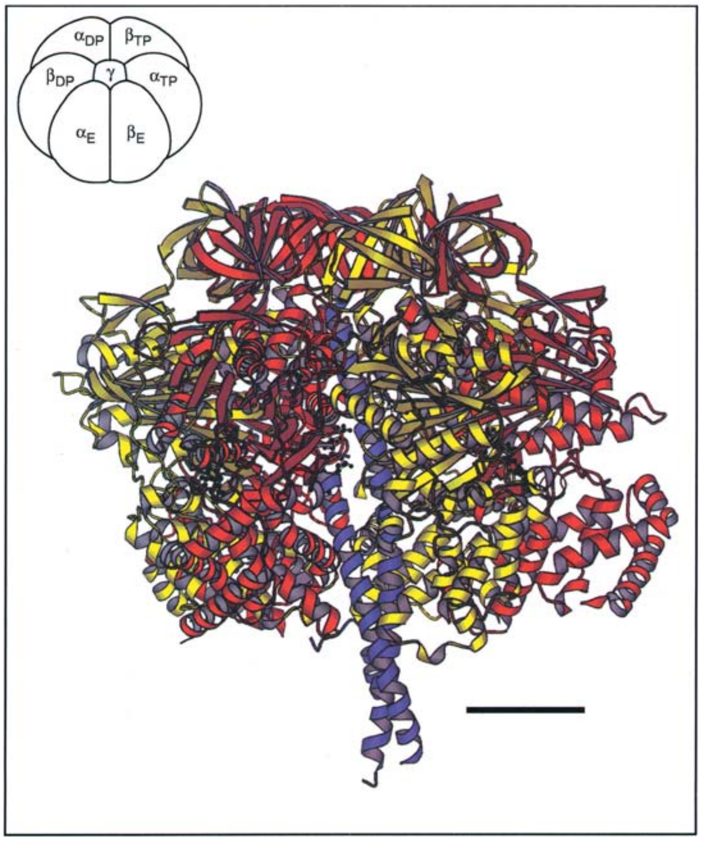


Figure 8.1.8.3
The protein crystal structure of F$_1$ ATPase, one of the largest non-symmetrical protein structure complexes, solved using SR data recorded on an image plate at the SRS wiggler 9.6, Daresbury. The scale bar is 20 Å long. Reprinted with permission from *Nature* (Abrahams *et al.*, 1994). Copyright (1994) MacMillan Magazines Limited.

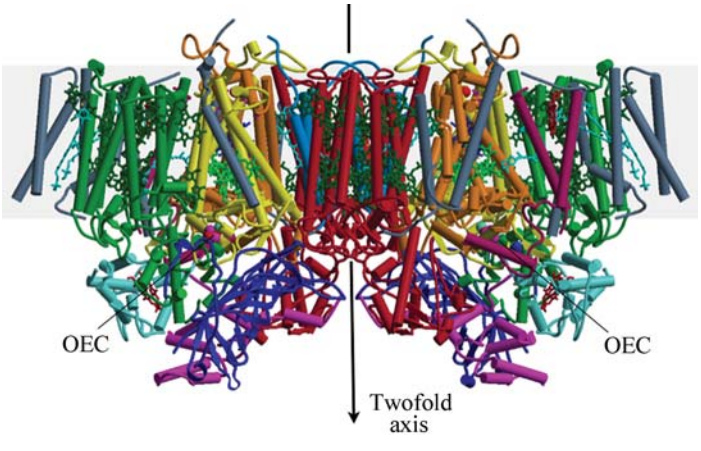


Figure 8.1.8.4
Side view of the structure of Photosystem II, the water-splitting enzyme of photosynthesis, determined using X-ray crystallography based on data recorded at the SLS and ESRF (Ferreira *et al.*, 2004). Kindly provided by Professor So Iwata, Imperial College, London.

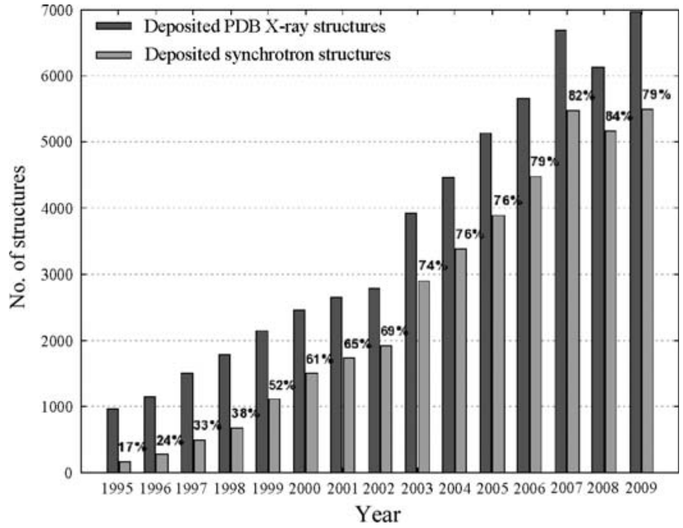


Figure 8.1.9.1

Synchrotron structures deposited in the PDB *versus* all PDB deposited structures (from http://biosync.rcsb.org/BiosyncStat.html) as of December 2009, *i.e.* a majority of all determined macromolecular crystal structures are now synchrotron-radiation derived.

discovery [for an early description see Bugg *et al.* (1993), for an example of pharmaceutical company collaboration see the Industrial Macromolecular Crystallography Association (IMCA) Collaborative Access Team at APS in Chicago (http://www.imca.aps.anl.gov/) and for a recent European perspective see *e.g.* Maclean *et al.* (2006).]

8.1.8.6. Radiation damage

The successes of these sources in macromolecular crystallography have been spectacular, so much so that provision of such beamlines now accounts for ~40% of the APS and ESRF insertion-device sectors. Nevertheless, X-radiation damage has been a continuing concern not least as the 'cutting edge' of capability to study ever-smaller crystals and ever-larger unit cells at synchrotron-radiation sources has been improved and optimized. A notable development has been a series of International Workshops on X-ray Damage to Crystalline Biological Samples. Most of these workshops have resulted in special issues of the *Journal of Synchrotron Radiation* (Volume 9 part 6, Volume 12 part 3, Volume 14 part 1 and Volume 16 part 2). The latest of these has a mini-review by Garman & Nave (2009).

8.1.9. Concluding remarks

SR and crystallography are now intricately intertwined in their scientific futures and in facilities provision (see *e.g.* Helliwell, 1998; Dauter, 2006; Fig. 8.1.9.1). Jiang & Sweet (2004) have given a systematic analysis of the impact of SR on macromolecular crystallography capabilities.

The new XFELs have produced very novel possibilities for 3D structure determination not only of non-crystallizable proteins but also whole cells (Neutze *et al.*, 2000; Miao *et al.*, 2001; Sayre, 2008; Shapiro, 2008). Whilst these approaches have attracted controversy, they represent a bold new push of X-ray diffraction methods towards widening capabilities in important frontiers in structural biology and structural cellular biology.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F_1 ATPase. Acta Cryst.* D**52**, 30–42.

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Structure at 2.8 Å resolution of F1-ATPase from bovine heart mitochondria. Nature (London)*, **370**, 621–628.

Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D. & Brown, F. (1989). *The 3-dimensional structure of foot and mouth disease virus at 2.9 Å resolution. Nature (London)*, **337**, 709–716.

Allinson, N. M. (1994). *Development of non-intensified charge-coupled device area X-ray detectors. J. Synchrotron Rad.* **1**, 54–62.

Amemiya, Y. (1997). *X-ray storage phosphor imaging plate detectors: high sensitivity X-ray area detector. Methods Enzymol.* **276**, 233–243.

Andrews, S. J., Hails, J. E., Harding, M. M. & Cruickshank, D. W. J. (1987). *The mosaic spread of very small crystals deduced from Laue diffraction patterns. Acta Cryst.* A**43**, 70–73.

Arndt, U. W., Greenhough, T. J., Helliwell, J. R., Howard, J. A. K., Rule, S. A. & Thompson, A. W. (1982). *Optimised anomalous dispersion crystallography: a synchrotron X-ray polychromatic simultaneous profile method. Nature (London)*, **298**, 835–838.

Arnold, E., Vriend, G., Luo, M., Griffith, J. P., Kamer, G., Erickson, J. W., Johnson, J. E. & Rossmann, M. G. (1987). *The structure determination of a common cold virus, human rhinovirus 14. Acta Cryst.* A**43**, 346–361.

Arzt, S., Campbell, J. W., Harding, M. M., Hao, Q. & Helliwell, J. R. (1999). *LSCALE – the new normalization, scaling and absorption correction program in the Daresbury Laue software suite. J. Appl. Cryst.* **32**, 554–562.

Baker, P. J., Farrants, G. W., Stillman, T. J., Britton, K. L., Helliwell, J. R. & Rice, D. W. (1990). *Isomorphous replacement with optimized anomalous scattering applied to protein crystallography. Acta Cryst.* A**46**, 721–725.

Ban, N., Freeborn, B., Nissen, P., Penczek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. & Steitz, T. A. (1998). *A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. Cell*, **93**, 1105–1115.

Barber, J. (2009). *Photosynthetic energy conversion: natural and artificial. Chem. Soc. Rev.* **38**, 185–196.

Bartunik, H. D., Clout, P. N. & Robrahn, B. (1981). *Rotation data collection for protein crystallography with time-variable incident intensity from synchrotron radiation sources. J. Appl. Cryst.* **14**, 134–136.

Beuville, E., Beche, J. F., Cork, C., Douence, V., Earnest, T., Millaud, J., Nygren, D., Padmore, H., Turko, B., Zizka, G., Datte, P. & Xuong, N. H. (1997). *A 16 × 16 pixel array detector for protein crystallography. Nucl. Instrum. Methods*, **395**, 429–434.

Bilderback, D. H. (1986). *The potential of cryogenic silicon and germanium X-ray monochromators for use with large synchrotron heat loads. Nucl. Instrum. Methods*, **246**, 434–436.

Blakeley, M. P. (2009). *Neutron macromolecular crystallography. Crystallogr. Rev.* **15**, 157–218.

Blewett, J. P. (1946). *Radiation losses in the induction electron accelerator. Phys. Rev.* **69**, 87–95.

Blow, D. M. (2002). *Rearrangement of Cruickshank's formulae for the diffraction-component precision index. Acta Cryst.* D**58**, 792–797.

Bonse, U., Materlik, G. & Schröder, W. (1976). *Perfect-crystal monochromators for synchrotron X-radiation. J. Appl. Cryst.* **9**, 223–230.

Bourenkov, G. P., Popov, A. N. & Bartunik, H. D. (1996). *A Bayesian approach to Laue diffraction analysis and its potential for time-resolved protein crystallography. Acta Cryst.* A**52**, 797–811.

Bourgeois, D. & Weik, M. (2009). *Kinetic protein crystallography: a tool to watch proteins in action. Crystallogr. Rev.* **15**, 87–118.

Brammer, R., Helliwell, J. R., Lamb, W., Liljas, A., Moore, P. R., Thompson, A. W. & Rathbone, K. (1988). *A new protein crystallography station on the SRS wiggler beamline for very rapid Laue and rapidly tunable monochromatic experiments: I. Design principles, ray tracing and heat calculations. Nucl. Instrum. Methods A*, **271**, 678–687.

Branden, C. I. (1994). *The new generation of synchrotron machines. Structure*, **2**, 5–6.

Brinkmann, R., Materlik, G., Rossbach, J., Schneider, J. R. & Wilk, B. H. (1997). *An X-ray FEL laboratory as part of a linear collider design. Nucl. Instrum. Methods*, **393**, No. 1–3, 86–92.

Broennimann, Ch., Eikenberry, E. F., Henrich, B., Horisberger, R., Huelsen, G., Pohl, E., Schmitt, B., Schulze-Briese, C., Suzuki, M., Tomizaki, T., Toyokawa, H. & Wagner, A. (2006). *The PILATUS 1M detector J. Synchrotron Rad.* **13**, 120–130.

Bugg, C. E., Carson, W. M. & Montgomery, J. A. (1993). *Drugs by design. Sci. Am.* **269**, 60–66.

Campbell, J. W. & Hao, Q. (1993). *Evaluation of reflection intensities for the components of multiple Laue diffraction spots. II. Using the wavelength-normalization curve. Acta Cryst.* A**49**, 889–893.

Caspar, D. L. D., Clarage, J., Salunke, D. M. & Clarage, M. (1988). *Liquid-like movements in crystalline insulin. Nature (London)*, **352**, 659–662.

Cassetta, A., Deacon, A., Emmerich, C., Habash, J., Helliwell, J. R., McSweeney, S., Snell, E., Thompson, A. W. & Weisgerber, S. (1993). *The emergence of the synchrotron Laue method for rapid data collection from protein crystals. Proc. R. Soc. London Ser. A*, **442**, 177–192.

Cauchois, Y., Bonnelle, C. & Missoui, G. (1963). *Rayonnement electromagnetique – premiers spectres X du rayonnement d'orbite du synchrotron de Frascati. C. R. Acad. Sci. Paris*, **257**, 409–412.

Charpak, G. (1970). *Evolution of the automatic spark chambers. Annu. Rev. Nucl. Sci.* **20**, 195–254.

Chayen, N. E., Boggon, T. J., Cassetta, A., Deacon, A., Gleichmann, T., Habash, J., Harrop, S. J., Helliwell, J. R., Nieh, Y. P., Peterson, M. R., Raftery, J., Snell, E. H., Hädener, A., Niemann, A. C., Siddons, D. P., Stojanoff, V., Thompson, A. W., Ursby, T. & Wulff, M. (1996). *Trends and challenges in experimental macromolecular crystallography. Q. Rev. Biophys.* **29**, 227–278.

Chayen, N. E. & Helliwell, J. R. (1998). *Protein crystallography: the human genome in 3-D. Phys. World*, **11**, No. 5, 43–48.

Cianci, M., Antonyuk, S., Bliss, N., Bailey, M. W., Buffey, S. G., Cheung, K. C., Clarke, J. A., Derbyshire, G. E., Ellis, M. J., Enderby, M. J., Grant, A. F., Holbourn, M. P., Laundy, D., Nave, C., Ryder, R., Stephenson, P., Helliwell, J. R. & Hasnain, S. S. (2005). *A high-throughput structural biology/proteomics beamline at the SRS on a new multipole wiggler. J. Synchrotron Rad.* **12**, 455–466.

Cianci, M., Helliwell, J. R. & Suzuki, A. (2008). *The interdependence of wavelength, redundancy and dose in sulfur SAD experiments. Acta Cryst.* D**64**, 1196–1209.

Cianci, M., Rizkallah, P. J., Olczak, A., Raftery, J., Chayen, N. E., Zagalsky, P. F. & Helliwell, J. R. (2001). *Structure of lobster apocrustacyanin* A_1 *using softer X-rays. Acta Cryst.* D**57**, 1219–1229.

Colapietro, M., Cappuccio, G., Marciante, C., Pifferi, A., Spagna, R. & Helliwell, J. R. (1992). *The X-ray diffraction station at the ADONE wiggler facility: preliminary results (including crystal perfection). J. Appl. Cryst.* **25**, 192–194.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J., Helliwell, J. R. & Johnson, L. N. (1992). Editors. *Time-resolved Macromolecular Crystallography.* Oxford: OUP/The Royal Society.

Cruickshank, D. W. J., Helliwell, J. R. & Moffat, K. (1987). *Multiplicity distribution of reflections in Laue diffraction. Acta Cryst.* A**43**, 656–674.

Cruickshank, D. W. J., Helliwell, J. R. & Moffat, K. (1991). *Angular distribution of reflections in Laue diffraction. Acta Cryst.* A**47**, 352–373.

Dauter, Z. (2006). *Current state and prospects of macromolecular crystallography. Acta Cryst.* (2006). D**62**, 1–11.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution. Curr. Opin. Struct. Biol.* **7**, 681–688.

Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *The structure of concanavalin A and its bound solvent determined with small-molecule accuracy at 0.94 Å resolution. J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.

Doniach, S., Hodgson, K., Lindau, I., Pianetta, P. & Winick, H. (1997). *Early work with synchrotron radiation at Stanford. J. Synchrotron Rad.* **4**, 380–395.

Doucet, J. & Benoit, J. P. (1987). *Molecular dynamics by analysis of the X-ray diffuse scattering from lysozyme crystals. Nature (London)*, **325**, 643–646.

Dumas, C., Duquerroy, S. & Janin, J. (1995). *Phasing with mercury at 1 Å wavelength. Acta Cryst.* D**51**, 814–818.

Eikenberry, E. F., Barna, S. L., Tate, M. W., Rossi, G., Wixted, R. L., Sellin, P. J. & Gruner, S. M. (1998). *A pixel-array detector for time-resolved X-ray diffraction. J. Synchrotron. Rad.* **5**, 252–255.

Einspahr, H., Suguna, K., Suddath, F. L., Ellis, G., Helliwell, J. R. & Papiz, M. Z. (1985). *The location of manganese and calcium ion cofactors in pea lectin crystals by use of anomalous dispersion and tuneable synchrotron X-radiation. Acta Cryst.* B**41**, 336–341.

Elleaume, P. (1989). *Special synchrotron radiation sources. Part 1: conventional insertion devices. Synchrotron Radiat. News*, **1**, 18–23.

Elleaume, P. (1998). *Two-plane focusing of 30 keV undulator radiation. J. Synchrotron Rad.* **5**, 1–5.

ESRF Foundation Phase Report (1987). Grenoble: ESRF.

Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. (2004). *Architecture of the photosynthetic oxygen-evolving center. Science*, **303**, 1831–1838.

Fourme, R. (1997). *Position-sensitive gas detectors: MWPCs and their gifted descendants. Nucl. Instrum. Methods A*, **392**, 1–11.

Fourme, R., Kahn, R., Mezouar, M., Girard, E., Hoerentrup, C., Prangé, T. & Ascone, I. (2001). *High-pressure protein crystallography (HPPX): instrumentation, methodology and results on lysozyme crystals. J. Synchrotron Rad.* **8**, 1149–1156.

Freund, A. K. (1996). *Third-generation synchrotron radiation X-ray optics. Structure*, **4**, 121–125.

Garman, E. F. & Nave, C. (2009). *Radiation damage in protein crystals examined under various conditions by different methods J. Synchrotron Rad.* **16**, 129–132.

Girard, E., Dhaussy, A.-C., Couzinet, B., Chervin, J.-C., Mezouar, M., Kahn, R., Ascone, I. & Fourme, R. (2007). *Toward fully fledged high-pressure macromolecular crystallography. J. Appl. Cryst.* **40**, 912–918.

Glover, I. D., Harris, G. W., Helliwell, J. R. & Moss, D. S. (1991). *The variety of X-ray diffuse scattering from macromolecular crystals and its respective components. Acta Cryst.* B**47**, 960–968.

Gonzalez, A., Cohen, A., Eriksson, T., McPhillips, S., Moorhead, P., Narevicius, J., Sharp, K., Smith, C., Song, J. & Soltis, S. M. (2005). *Facility updates: remote access to the SSRL macromolecular crystallography beamlines. Synchrotron Radiat. News*, **18**, 36–39.

Greenhough, T. J. & Helliwell, J. R. (1982). *Oscillation camera data processing: reflecting range and prediction of partiality. II. Monochromatised synchrotron X-radiation from a singly bent triangular monochromator. J. Appl. Cryst.* **15**, 493–508.

Grimes, J. M., Burroughs, J. N., Gouet, P., Diprose, J. M., Malby, R., Zientara, S., Mertens, P. P. C. & Stuart, D. I. (1998). *The atomic structure of the bluetongue virus core. Nature (London)*, **395**, 470–478.

Grimes, J. M., Jakana, J., Ghosh, M., Basak, A. K., Roy, P., Chiu, W., Stuart, D. I. & Prasad, B. V. V. (1997). *An atomic model of the outer layer of the bluetongue virus core derived from X-ray crystallography and electron cryomicroscopy. Structure*, **5**, 885–893.

Gruner, S. M. & Ealick, S. E. (1995). *Charge coupled device X-ray detectors for macromolecular crystallography. Structure*, **3**, 13–15.

Guillot, B., Jelsch, C., Podjarny, A. & Lecomte, C. (2008). *Charge-density analysis of a protein structure at subatomic resolution: the human aldose reductase case. Acta Cryst.* D**64**, 567–588.

Guss, J. M., Merritt, E. A., Phizackerley, R. P., Hedman, B., Murata, M., Hodgson, K. O. & Freeman, H. C. (1988). *Phase determination by multiple wavelength X-ray diffraction – crystal structure of a basic blue copper protein from cucumbers. Science*, **241**, 806–811.

Habash, J., Raftery, J., Nuttall, R., Price, H. J., Wilkinson, C., Kalb (Gilboa), A. J. & Helliwell, J. R. (2000). *Direct determination of the positions of the deuterium atoms of the bound water in concanavalin A by neutron Laue crystallography. Acta Cryst.* D**56**, 541–550.

Habash, J., Raftery, J., Weisgerber, S., Cassetta, A., Lehmann, M. S., Hoghoj, P., Wilkinson, C., Campbell, J. W. & Helliwell, J. R. (1997). *Neutron Laue diffraction study of concanavalin A: the proton of Asp28. J. Chem. Soc. Faraday Trans.* **93**, 4313–4317.

Hädener, A., Matzinger, P. K., Battersby, A. R., McSweeney, S., Thompson, A. W., Hammersley, A. P., Harrop, S. J., Cassetta, A., Deacon, A., Hunter, W. N., Nieh, Y. P., Raftery, J., Hunter, N. & Helliwell, J. R. (1999). *Determination of the structure of seleno-methionine-labelled hydroxymethylbilane synthase in its active form by multi-wavelength anomalous dispersion. Acta Cryst.* D**55**, 631–643.

Hajdu, J., Acharya, K. R., Stuart, D. I., McLaughlin, P. J., Barford, D., Oikonomakos, N. G., Klein, H. & Johnson, L. N. (1987). *Catalysis in the crystal – synchrotron radiation studies with glycogen phosphorylase. EMBO J.* **6**, 539–546.

Hajdu, J., Machin, P. A., Campbell, J. W., Greenhough, T. J., Clifton, I. J., Zurek, S., Gover, S., Johnson, L. N. & Elder, M. (1987). *Millisecond X-ray diffraction and the first electron density map from Laue photographs of a protein crystal. Nature (London)*, **329**, 178–181.

Hall, G. (1995). *Silicon pixel detectors for X-ray diffraction studies at synchrotron sources. Q. Rev. Biophys.* **28**, 1–32.

Harmsen, A., Leberman, R. & Schultz, G. E. (1976). *Comparison of protein crystal diffraction patterns and absolute intensities from synchrotron and conventional X-ray sources. J. Mol. Biol.* **104**, 311–314.

Hart, M. (1971). *Bragg reflection X-ray optics. Rep. Prog. Phys.* **34**, 435–490.

Haslegrove, J. C., Faruqi, A. R., Huxley, H. E. & Arndt, U. W. (1977). *The design and use of a camera for low-angle X-ray diffraction experiments with synchrotron radiation. J. Phys. E*, **10**, 1035–1044.

Hastings, J. B. (1977). *X-ray optics and monochromators for synchrotron radiation. J. Appl. Phys.* **48**, 1576–1584.

Hastings, J. B., Kincaid, B. M. & Eisenberger, P. (1978). *A separated function focusing monochromator system for synchrotron radiation. Nucl. Instrum. Methods*, **152**, 167–171.

Hedman, B., Hodgson, K. O., Helliwell, J. R., Liddington, R. & Papiz, M. Z. (1985). *Protein micro-crystal diffraction and the effects of radiation damage with ultra high flux synchrotron radiation. Proc. Natl Acad. Sci. USA*, **82**, 7604–7607.

Helliwell, J. R. (1977). *Application of synchrotron radiation to protein crystallography: preliminary experiments on 6PDGH crystals using the NINA synchrotron, Daresbury, UK.* DPhil thesis, University of Oxford, Appendix 1.

Helliwell, J. R. (1979). *Optimisation of anomalous scattering and structural studies of proteins using synchrotron radiation.* In *Proceedings of the Study Weekend. Applications of SR to the Study of Large Molecules of Chemical and Biological Interest*, edited by R. B. Cundall & I. H. Munro, pp. 1–6. DL/SCI/R13. Warrington: Daresbury Laboratory.

Helliwell, J. R. (1984). *Synchrotron X-radiation protein crystallography: instrumentation, methods and applications. Rep. Prog. Phys.* **47**, 1403–1497.

Helliwell, J. R. (1985). *Protein crystallography with synchrotron radiation. J. Mol. Struct.* **130**, 63–91.

Helliwell, J. R. (1988). *Protein crystal perfection and the nature of radiation damage. J. Crystal Growth*, **90**, 259–272.

Helliwell, J. R. (1992). *Macromolecular Crystallography with Synchrotron Radiation.* Cambridge University Press.

Helliwell, J. R. (1993). *The choice of X-ray wavelength in macromolecular crystallography*. In *Computational Aspects of Data Collection*, compiled by S. Bailey, pp. 80–88. DL/SCI/R34. Warrington: Daresbury Laboratory.

Helliwell, J. R. (1997*a*). *Overview on synchrotron radiation and application in macromolecular crystallography. Methods Enzymol.* **276**, 203–217.

Helliwell, J. R. (1997*b*). *Neutron Laue diffraction does it faster. Nature Struct. Biol.* **4**, 874–876.

Helliwell, J. R. (1998). *Synchrotron radiation and crystallography: the first fifty years. Acta Cryst.* A**54**, 738–749.

Helliwell, J. R. (2004). *Overview and new developments in softer X-ray* ($2\ \text{Å} < \lambda < 5\ \text{Å}$) *protein crystallography*. *J. Synchrotron Rad.* **11**, 1–3.

Helliwell, J. R., Ealick, S., Doing, P., Irving, T. & Szebenyi, M. (1993). *Towards the measurement of ideal data for macromolecular crystallography using synchrotron sources. Acta Cryst.* D**49**, 120–128.

Helliwell, J. R., Habash, J., Cruickshank, D. W. J., Harding, M. M., Greenhough, T. J., Campbell, J. W., Clifton, I. J., Elder, M., Machin, P. A., Papiz, M. Z. & Zurek, S. (1989). *The recording and analysis of synchrotron X-radiation Laue diffraction photographs. J. Appl. Cryst.* **22**, 483–497.

Helliwell, J. R., Nieh, Y. P., Raftery, J., Cassetta, A., Habash, J., Carr, P. D., Ursby, T., Wulff, M., Thompson, A. W., Niemann, A. C. & Hädener, A. (1998). *Time-resolved structures of hydroxymethylbilane synthase (Lys59Gln mutant) as it is loaded with substrate in the crystal determined by Laue diffraction. J. Chem. Soc. Faraday Trans.* **94**, 2615–2622.

Helliwell, J. R., Papiz, M. Z., Glover, I. D., Habash, J., Thompson, A. W., Moore, P. R., Harris, N., Croft, D. & Pantos, E. (1986). *The wiggler protein crystallography workstation at the Daresbury SRS; progress and results. Nucl. Instrum. Methods A*, **246**, 617–623.

Helliwell, J. R. & Rentzepis, P. M. (1997). Editors. *Time-resolved Diffraction.* Oxford University Press.

Helliwell, J. R. & Wilkinson, C. (1994). *X-ray and neutron Laue diffraction*. In *Neutron and Synchrotron Radiation for Condensed Matter Studies*, Vol. 3, edited by J. Baruchel, J. L. Hodeau, M. S. Lehmann, J. R. Regnard & C. Schlenker, ch. 12. Berlin: Springer Verlag.

Hendrickson, W. A. (1985). *Analysis of protein structure from diffraction measurements at multiple wavelengths. Trans. Am. Crystallogr. Assoc.* **21**, 11–21.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multi-wavelength anomalous diffraction (MAD) – a vehicle for direct determination of 3-dimensional structure. EMBO J.* **9**, 1665–1672.

Hendrickson, W. A., Pahler, A., Smith, J. L., Satow, Y., Merritt, E. A. & Phizackerley, R. P. (1989). *Crystal structure of core streptavadin determined from multi-wavelength anomalous diffraction of synchrotron radiation. Proc. Natl Acad. Sci. USA*, **86**, 2190–2194.

Herzenberg, A. & Lau, H. S. M. (1967). *Anomalous scattering and the phase problem. Acta Cryst.* **22**, 24–28.

Holmes, K. C. (1998). *A molecular model for muscle contraction. Acta Cryst.* A**54**, 789–797.

Hope, H., Frolow, F., von Böhlen, K., Makowski, I., Kratky, C., Halfon, Y., Danz, H., Webster, P., Bartels, K. S., Wittmann, H. G. & Yonath, A. (1989). *Cryocrystallography of ribosomal particles. Acta Cryst.* B**45**, 190–199.

Hoppe, W. & Jakubowski, V. (1975). *The determination of phases of erythrocruorin using the two-wavelength method with iron as anomalous scatterer.* In *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams, pp. 437–461. Copenhagen: Munksgaard.

Huxley, H. E. & Holmes, K. C. (1997). *Development of synchrotron radiation as a high-intensity source for X-ray diffraction. J. Synchrotron Rad.* **4**, 366–379.

Iwanenko, D. & Pomeranchuk, I. (1944). *On the maximal energy attainable in a betatron. Phys. Rev.* **65**, 343.

Jakoncic, J., Di Michiel, M., Zhong, Z., Honkimaki, V., Jouanneau, Y. & Stojanoff, V. (2006). *Anomalous diffraction at ultra-high energy for protein crystallography. J. Appl. Cryst.* **39**, 831–841.

Jiang, J. & Sweet, R. M. (2004). *Protein Data Bank depositions from synchrotron sources*. *J. Synchrotron Rad.* **11**, 319–327.

Kahn, R., Fourme, R., Bosshard, R., Chiadmi, M., Risler, J. L., Dideberg, O. & Wery, J. P. (1985). *Crystal structure study of opsanus-tau parvalbumin by multi-wavelength anomalous dispersion. FEBS Lett.* **170**, 133–137.

Karle, J. (1967). *Anomalous scatterers in X-ray diffraction and the use of several wavelengths. Appl. Opt.* **6**, 2132–2135.

Karle, J. (1980). *Some anomalous dispersion developments for the structure investigation of macromolecular systems in biology. Int. J. Quantum Chem. Symp.* **7**, 357–367.

Karle, J. (1989). *Macromolecular structure from anomalous dispersion. Phys. Today*, **42**, 22–29.

Karle, J. (1994). *Developments in anomalous scattering for structure determination.* In *Resonant Anomalous X-ray Scattering*, edited by G. Materlik, C. J. Sparks & K. Fischer, pp. 145–158. Amsterdam: North Holland.

Kohra, K., Ando, M., Matsushita, T. & Hashizume, H. (1978). *Design of high resolution X-ray optical system using dynamical diffraction for synchrotron radiation. Nucl. Instrum. Methods*, **152**, 161–166.

Korszun, Z. R. (1987). *The tertiary structure of azurin from pseudomonas-denitrificans as determined by Cu resonant diffraction using synchrotron radiation. J. Mol. Biol.* **196**, 413–419.

Lairson, B. M. & Bilderback, D. H. (1982). *Transmission X-ray mirror – a new optical element. Nucl. Instrum. Methods*, **195**, 79–83.

Lemonnier, M., Fourme, R., Rousseaux, F. & Kahn, R. (1978). *X-ray curved-crystal monochromator system at the storage ring DCI. Nucl. Instrum. Methods*, **152**, 173–177.

Lewis, R. (1994). *Multiwire gas proportional counters: decrepit antiques or classic performers? J. Synchrotron Rad.* **1**, 43–53.

Liddington, R. C., Yan, Y., Moulai, J., Sahli, R., Benjamin, T. L. & Harrison, S. C. (1991). *Structure of simian virus-40 at 3.8 Å resolution. Nature (London)*, **354**, 278–284.

Liljas, A., Liljas, L., Piskur, J., Lindblom, G., Nissen, P. & Kjeldgaard, M. (2009). *Textbook of Structural Biology*. Singapore: World Scientific.

Lindley, P. F. (1999). *Macromolecular crystallography with a third-generation synchrotron source. Acta Cryst.* D**55**, 1654–1662.

Liu, Y., Ogata, C. & Hendrickson, W. (2001). *Multiwavelength anomalous diffraction analysis at the M absorption edges of uranium. Proc. Natl Acad. Sci. USA*, **98**, 10648–10653.

Lovelace, J., Snell, E. H., Pokross, M., Arvai, A. S., Nielsen, C., Xuong, N.-H., Bellamy, H. D. & Borgstahl, G. E. O. (2000). *BEAM-ish: a graphical user interface for the physical characterization of macromolecular crystals. J. Appl. Cryst.* **33**, 1187–1188.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). *Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature (London)*, **389**, 251–260.

Luger, P. (2007). *Fast electron density methods in the life sciences – a routine application in the future? Org. Biomol. Chem.* **5**, 2529–2540.

McDermott, G., Prince, S. M., Freer, A. A., Hawthornthwaite-Lawless, A. M., Papiz, M. Z., Cogdell, R. J. & Isaacs, N. W. (1995). *Crystal structure of an integral membrane light harvesting complex from photosynthetic bacteria. Nature (London)*, **374**, 517–521.

Maclean, E. J., Rizkallah, P. J. & Helliwell, J. R. (2006). *Protein crystallography and synchrotron radiation; current status and future landscape. Eur. Pharm. Rev.* **2**, 71–76.

Mayans, O. & Wilmanns, M. (1999). *X-ray analysis of protein crystals with thin-plate morphology. J. Synchrotron Rad.* **6**, 1016–1020.

Miao, J., Hodgson, K. O. & Sayre, D. (2001). *An approach to three-dimensional structures of biomolecules by using single-molecule diffraction images. Proc. Natl Acad. Sci. USA*, **98**, 6641–6645.

Miller, A. (1994). *Advanced Synchrotron Sources – Plans at ESRF in SR in Biophysics*, edited by S. S. Hasnain. Chichester: Ellis Horwood.

Mills, D. M., Helliwell, J. R., Kvick, Å., Ohta, T., Robinson, I. A. & Authier, A. (2005). *Report of the Working Group on Synchrotron Radiation Nomenclature – brightness, spectral brightness or brilliance? J. Synchrotron Rad.* **12**, 385.

Moffat, K., Szebenyi, D. & Bilderback, D. H. (1984). *X-ray Laue diffraction from protein crystals. Science*, **223**, 1423–1425.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *The use of MULTAN to locate the positions of anomalous scatterers. Acta Cryst.* A**45**, 715–718.

Munro, I. H. (1997). *Synchrotron radiation research in the UK. J. Synchrotron Rad.* **4**, 344–358.

Neutze, R. & Hajdu, J. (1997). *Femtosecond time resolution in X-ray diffraction experiments. Proc. Natl Acad. Sci. USA*, **94**, 5651–5655.

Neutze, R., Wouts, R., Spoel, D., Weckert, E. & Hajdu, J. (2000). *Potential for biomolecular imaging with femtosecond X-ray pulses. Nature (London)*, **406**, 752–757.

Nieh, Y. P., Raftery, J., Weisgerber, S., Habash, J., Schotte, F., Ursby, T., Wulff, M., Hädener, A., Campbell, J. W., Hao, Q. & Helliwell, J. R. (1999). *Accurate and highly complete synchrotron protein crystal Laue diffraction data using the ESRF CCD and the Daresbury Laue software. J. Synchrotron Rad.* **6**, 995–1006.

Niemann, A. C., Matzinger, P. K. & Hadener, A. (1994). *A kinetic analysis of the reaction catalysed by hydroxymethylbilane synthase. Helv. Chim. Acta*, **77**, 1791–1809.

Noller, H. F. (2005). *RNA structure: reading the ribosome. Science*, **309**, 1508–1514.

Nugent, K. (2009). *Coherent methods in the X-ray sciences. Adv. Phys.* **59**, 1–99.

Okaya, Y. & Pepinsky, R. (1956). *New formulation and solution of the phase problem in X-ray analysis of non-centric crystals containing anomalous scatterers. Phys. Rev.* **103**, 1645–1647.

Parratt, L. G. (1959). *Use of synchrotron orbit-radiation in X-ray physics. Rev. Sci. Instrum.* **30**, 297–299.

Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. (1997). *X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipidic cubic phases. Science*, **277**, 1676–1681.

Perrakis, A., Cipriani, F., Castagna, J.-C., Claustre, L., Burghammer, M., Riekel, C. & Cusack, S. (1999). *Protein microcrystals and the design of a microdiffractometer: current experience and plans at EMBL and ESRF/ID13. Acta Cryst.* D**55**, 1765–1770.

Peterson, M. R., Harrop, S. J., McSweeney, S. M., Leonard, G. A., Thompson, A. W., Hunter, W. N. & Helliwell, J. R. (1996). *MAD phasing strategies explored with a brominated oligonucleotide crystal at 1.65 Å resolution. J. Synchrotron Rad.* **3**, 24–34.

Phillips, J. C. & Hodgson, K. O. (1980). *The use of anomalous scattering effects to phase diffraction patterns from macromolecules. Acta Cryst.* A**36**, 856–864.

Phillips, J. C., Templeton, L. K., Templeton, D. H. & Hodgson, K. O. (1978). *L-III edge anomalous X-ray scattering by cesium measured with synchrotron radiation. Science*, **201**, 257–259.

Phillips, J. C., Wlodawer, A., Goodfellow, J. M., Watenpaugh, K. D., Sieker, L. C., Jensen, L. H. & Hodgson, K. O. (1977). *Applications of synchrotron radiation to protein crystallography. II. Anomalous scattering, absolute intensity and polarization. Acta Cryst.* A**33**, 445–455.

Phillips, J. C., Wlodawer, A., Yevitz, M. M. & Hodgson, K. O. (1976). *Applications of synchrotron radiation to protein crystallography: preliminary results. Proc. Natl Acad. Sci. USA*, **73**, 128–132.

Polikarpov, I., Teplyakov, A. & Oliva, G. (1997). *The ultimate wavelength for protein crystallography? Acta Cryst.* D**53**, 734–737.

Ren, Z. & Moffat, K. (1995*a*). *Quantitative analysis of synchrotron Laue diffraction patterns in macromolecular crystallography. J. Appl. Cryst.* **28**, 461–481.

Ren, Z. & Moffat, K. (1995*b*). *Deconvolution of energy overlaps in Laue diffraction. J. Appl. Cryst.* **28**, 482–493.

Riekel, C. (2000). *New avenues in X-ray microbeam experiments. Rep. Prog. Phys.* **63**, 233–262.

Riekel, C., Burghammer, M. & Schertler, G. (2005). *Protein crystallography microdiffraction. Curr. Opin. Struct. Biol.* **15**, 556–562.

Rosenbaum, G., Holmes, K. C. & Witz, J. (1971). *Synchrotron radiation as a source for X-ray diffraction. Nature (London)*, **230**, 434–437.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Structure of a human common cold virus and functional relationship to other picornaviruses. Nature (London)*, **317**, 145–153.

Rossmann, M. G. & Erickson, J. W. (1983). *Oscillation photography of radiation-sensitive crystals using a synchrotron source. J. Appl. Cryst.* **16**, 629–636.

Sakabe, N. (1983). *A focusing Weissenberg camera with multi-layer-line screens for macromolecular crystallography. J. Appl. Cryst.* **16**, 542–547.

Sakabe, N. (1991). *X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron radiation. Nucl. Instrum. Methods A*, **303**, 448–463.

Sakabe, N., Ikemizu, S., Sakabe, K., Higashi, T., Nagakawa, A. & Watanabe, N. (1995). *Weissenberg camera for macromolecules with imaging plate data collection at the Photon Factory – present status and future plan. Rev. Sci. Instrum.* **66**, 1276–1281.

Sayre, D. (2008). *Report on a project on three-dimensional imaging of the biological cell by single-particle X-ray diffraction. Acta Cryst.* A**64**, 33–35.

Schiltz, M. & Bricogne, G. (2009). *Instrument-independent specification of the diffraction geometry and polarization state of the incident X-ray beam. J. Appl. Cryst.* **42**, 101–108.

Schiltz, M., Kvick, Å., Svensson, O. S., Shepard, W., de La Fortelle, E., Prangé, T., Kahn, R., Bricogne, G. & Fourme, R. (1997). *Protein crystallography at ultra-short wavelengths: feasibility study of anomalous-dispersion experiments at the xenon K-edge. J. Synchrotron Rad.* **4**, 287–297.

Schwinger, J. (1949). *On the classical radiation of accelerated electrons. Phys. Rev.* **75**, 1912–1925.

Shapiro, D. A. (2008). *Report on a project on three-dimensional imaging of the biological cell by single-particle X-ray diffraction. Addendum. Acta Cryst.* A**64**, 36–37.

Snell, E. H., Weisgerber, S., Helliwell, J. R., Weckert, E., Hölzer, K. & Schroer, K. (1995). *Improvements in lysozyme protein crystal perfection through microgravity growth. Acta Cryst.* D**51**, 1099–1102.

Srajer, V., Teng, T. Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M. & Moffat, K. (1996). *Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography. Science*, **274**, 1726–1729.

Stuhrmann, H. B. & Lehmann, M. S. (1994). *Anomalous dispersion of X-ray scattering from low Z elements*. In *Resonant Anomalous X-ray Scattering*, edited by G. Materlik, C. J. Sparks & K. Fischer, pp. 175–194. Berlin: Springer Verlag.

Sugahara, M., Asada, Y., Shimada, H., Taka, H. & Kunishima, N. (2009). *HATODAS II – heavy-atom database system with potentiality scoring. J. Appl. Cryst.* **42**, 540–544.

Tate, M. W., Eikenberry, E. F., Barna, S. L., Wall, M. E., Lowrance, J. L. & Gruner, S. M. (1995). *A large-format high-resolution area X-ray detector based on a fiber-optically bonded charge-coupled device (CCD). J. Appl. Cryst.* **28**, 196–205.

Templeton, D. H. & Templeton, L. K. (1985). *X-ray dichroism and anomalous scattering of potassium tetrachloroplatinate(II). Acta Cryst.* A**41**, 365–371.

Templeton, D. H., Templeton, L. K., Phillips, J. C. & Hodgson, K. O. (1980). *Anomalous scattering of X-rays by cesium and cobalt measured with synchrotron radiation. Acta Cryst.* A**36**, 436–442.

Templeton, L. K., Templeton, D. H., Phizackerley, R. P. & Hodgson, K. O. (1982). *L_3-edge anomalous scattering by gadolinium and samarium measured at high resolution with synchrotron radiation. Acta Cryst.* A**38**, 74–78.

Teplyakov, A., Oliva, G. & Polikarpov, I. (1998). *On the choice of an optimal wavelength in macromolecular crystallography. Acta Cryst.* D**54**, 610–614.

Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. & Yoshikawa, S. (1996). *The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. Science*, **272**, 1136–1144.

Usha, R., Johnson, J. E., Moras, D., Thierry, J. C., Fourme, R. & Kahn, R. (1984). *Macromolecular crystallography with synchrotron radiation: collection and processing of data from crystals with a very large unit cell. J. Appl. Cryst.* **17**, 147–153.

Warren, J. E., Diakun, G., Bushnell-Wye, G., Fisher, S., Thalal, A., Helliwell, M. & Helliwell, J. R. (2008). *Science experiments via telepresence at a synchrotron radiation source facility*. *J. Synchrotron Rad.* **15**, 191–194.

Webb, N. G., Samson, S., Stroud, R. M., Gamble, R. C. & Baldeschwieler, J. D. (1977). *A focusing monochromator for small-angle diffraction studies with synchrotron radiation. J. Appl. Cryst.* **10**, 104–110.

Weckert, E. & Hümmer, K. (1997). *Multiple-beam X-ray diffraction for physical determination of reflection phases and its applications. Acta Cryst.* A**53**, 108–143.

Westbrook, E. M. & Naday, I. (1997). *Charge-coupled device-based area detectors. Methods Enzymol.* **276**, 244–268.

Wilson, K. S., Stura, E. A., Wild, D. L., Todd, R. J., Stuart, D. I., Babu, Y. S., Jenkins, J. A., Standing, T. S., Johnson, L. N., Fourme, R., Kahn, R., Gadet, A., Bartels, K. S. & Bartunik, H. D. (1983). *Macromolecular crystallography with synchrotron radiation. II. Results. J. Appl. Cryst.* **16**, 28–41.

Wimberley, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000). *Structure of the 30S ribosomal subunit*. *Nature (London)*, **407**, 327–339.

Winick, H. (1995). *The linac coherent light source (LCLS): a fourth-generation light source using the SLAC linac. J. Electron Spectrosc. Relat. Phenom.* **75**, 1–8.

Yano, J., Kern, J., Irrgang, K. D., Latimer, M. J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A. & Yachandra, V. K. (2005). *X-ray damage to the Mn4Ca complex in single crystals of photosystem II: a case study for metalloprotein crystallography. Proc. Natl Acad. Sci. USA*, **102**, 12047–12052.

Yonath, A. (1992). *Approaching atomic resolution in crystallography of ribosomes. Annu. Rev. Biophys. Biomol. Struct.* **21**, 77–93.

Yonath, A., Harms, J., Hansen, H. A. S., Bashan, A., Schlünzen, F., Levin, I., Koelln, I., Tocilj, A., Agmon, I., Peretz, M., Bartels, H., Bennett, W. S., Krumbholz, S., Janell, D., Weinstein, S., Auerbach, T., Avila, H., Piolleti, M., Morlang, S. & Franceschi, F. (1998). *Crystallographic studies on the ribosome, a large macromolecular assembly exhibiting severe nonisomorphism, extreme beam sensitivity and no internal symmetry. Acta Cryst.* A**54**, 945–955.

International Tables for Crystallography (2012). Vol. F, Chapter 8.2, pp. 205–210.

Chapter 8.2. Laue crystallography: time-resolved studies

K. Moffat

8.2.1. Introduction

The term 'Laue diffraction' describes the process of X-ray scattering that occurs when a stationary crystal is illuminated by a polychromatic beam of X-rays. The process therefore differs from the now more conventional diffraction techniques in which a moving crystal is illuminated by a monochromatic beam of X-rays. Although Laue diffraction was widely used for structure analysis in the early days of crystallography by Pauling, Bragg, Wyckoff and others, by the 1930s it was superseded by arguably simpler and more readily quantifiable monochromatic techniques. With the advent of naturally polychromatic synchrotron sources in the 1970s, it was natural to (re-)examine the suitability of the Laue technique. The first synchrotron-based Laue experiments to be published appear to have been those of Wood *et al.* (1983) for a small inorganic crystal and of Moffat *et al.* (1984) for a macromolecular crystal; see also Helliwell (1984, 1985). These experimenters realized that the Laue technique afforded exposure times that were short even with respect to those obtainable with similar crystals at the same synchrotron source using conventional monochromatic techniques, and much shorter than those obtainable with laboratory X-ray sources. This advantage, along with the use of a stationary crystal, the large number of Laue spots evident in a single image and the clear distinction of the Laue spots from the underlying X-ray background, suggested that the Laue technique might be particularly applicable to time-resolved crystallography. In this form of crystallography, the total X-ray scattering from the crystal (both the Bragg scattering and the diffuse, non-Bragg scattering) varies with time as the position and/or extent of order of the atoms in the crystal changes in response to some structural perturbation.

It is one thing to propose that a venerable technique may be applicable to a new class of experiments; it is quite another to identify and overcome the complexities and disadvantages of that technique, and to demonstrate how experiments should be conducted and raw data accurately reduced to structure amplitudes. It took roughly 15 years and the efforts of many investigators before it could be stated that Laue crystallography is coming of age (Ren *et al.*, 1999).

The redevelopment of Laue diffraction has depended on three main advances: the use of very intense polychromatic synchrotron sources; the realization that the so-called energy-overlap or overlapping-orders problem in Laue diffraction was theoretically tractable, of limited extent and could be overcome experimentally; and the development of appropriate algorithms and suitable software to address the energy-overlap, spatial-overlap and wavelength-normalization problems. All are discussed below.

Since both Laue crystallography and its applications to time-resolved studies have recently been described at length, this article emphasizes only the key points and directs the reader to the primary and review literature for the details.

8.2.2. Principles of Laue diffraction

The principles of Laue diffraction have been reviewed by Amorós *et al.* (1975), Cruickshank *et al.* (1987, 1991), Helliwell *et al.* (1989), Cassetta *et al.* (1993), Moffat (1997), and Ren *et al.* (1999).

Assume that a stationary, perfect single crystal that diffracts to a resolution limit of $d^*_{\max}$ is illuminated by a polychromatic X-ray beam spanning the wavelength (energy) range from $\lambda_{\min}$ ($E_{\max}$) to $\lambda_{\max}$ ($E_{\min}$). All reciprocal-lattice points that lie between the Ewald spheres of radii $1/\lambda_{\min}$ and $1/\lambda_{\max}$, and within a radius $d^*_{\max}$ of the origin O where $d^*_{\max} = 1/d_{\min}$, the resolution limit of the crystal, are in a diffracting position for a particular wavelength λ, where $\lambda_{\min} \leq \lambda \leq \lambda_{\max}$ and will contribute to a spot on the Laue diffraction pattern (Fig. 8.2.2.1). All such points diffract simultaneously and throughout the exposure, in contrast to a monochromatic diffraction pattern in which each point diffracts sequentially and briefly as it traverses the Ewald sphere. A Laue pattern may alternatively be thought of as the superposition of a series of monochromatic still patterns, each arising from a different wavelength in the range from $\lambda_{\min}$ to $\lambda_{\max}$.

Each Laue spot arises from the mapping of a complete ray (a central line in reciprocal space, emanating from the origin) onto a point on the detector. In contrast, each spot in a monochromatic pattern arises from the mapping of a single reciprocal-lattice point onto a point on the detector. A ray may contain only a single reciprocal-lattice point *hkl* with spacing d^*, in which case the corresponding Laue spot arises from a single wavelength (energy) and structure amplitude, or it may contain several

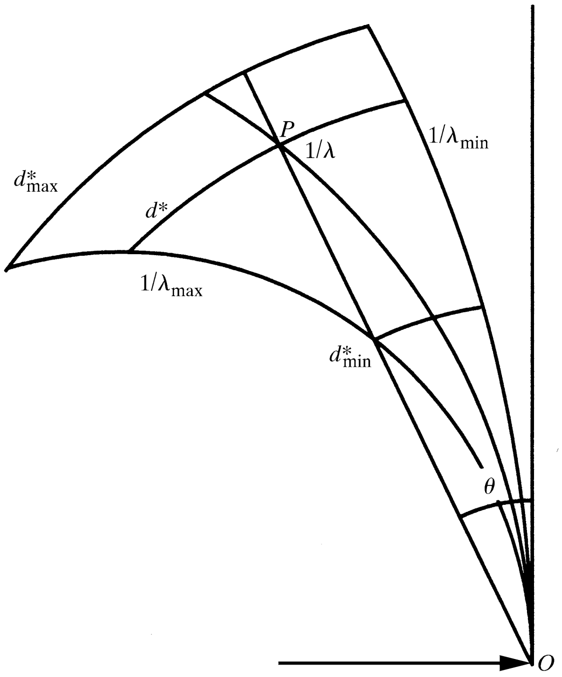


Figure 8.2.2.1
Laue diffraction geometry. The volume element dV stimulated in a Laue experiment lies between d^* and $d^* + dd^*$, between the Ewald spheres corresponding to λ and $\lambda + d\lambda$, and between φ and $\varphi + d\varphi$, where φ denotes rotation about the incident X-ray beam direction. The entire volume stimulated in a single Laue exposure lies between 0 and $d^*_{\max}$, between the Ewald spheres corresponding to $\lambda_{\min}$ and $\lambda_{\max}$, and between values of θ ranging from 0 to 2π.

Copyright © 2012 International Union of Crystallography

reciprocal-lattice points, such as $hkl, 2h2k2l \ldots nhnknl \ldots$, in which case the Laue spot contains several wavelengths (energies) and structure amplitudes. In the former case, the Laue spot is said to be single, and in the latter, multiple. The existence of multiple Laue spots is known as the energy-overlap problem: one spot contains contributions from several energies. It seems to have been thought by Pauling, Bragg and others that, as the wavelength range and the resolution limit $d^*_{\rm max}$ of the crystal increased, more and more Laue spots would be multiple and the energy-overlap problem would dominate. Cruickshank *et al.* (1987) showed that this was not so. Even in the extreme case of infinite wavelength range, no more than 12.5% of all Laue spots would be multiple. The energy-overlap problem is evidently of restricted extent. However, the magnitude of the energy-overlap problem varies with resolution: reciprocal-lattice points at low resolution are more likely to be associated with multiple Laue spots than to be single (Cruickshank *et al.*, 1987).

The extraction of X-ray structure amplitudes from a single Laue spot requires the derivation and application of a wavelength-dependent correction factor known as the wavelength normalization curve or λ-curve. This curve and other known factors relate the experimentally measured raw intensities of each Laue spot to the square of the corresponding structure amplitude. The integrated intensity of a Laue spot is achieved automatically by integration over wavelength, rather than in a monochromatic spot by integration over angle as the crystal rotates. If, however, a Laue spot is multiple, its total intensity arises from the sum of the integrated intensities of each of its components, known also as harmonics or orders *nhnknl* of the inner point *hkl* where *h*, *k* and *l* are co-prime.

Laue spots lie on conic sections, each corresponding to a central zone [*uvw*] in reciprocal space. Prominent spots known as nodal spots or nodals lie at the intersection of well populated zones and correspond to rays whose inner point *hkl* is of low co-prime indices. All nodal spots are multiple and all are surrounded by clear areas devoid of spots.

The volume of reciprocal space stimulated in a Laue exposure, V_v, is given by

$$V_v = 0.24\, d^{*4}_{\rm max}(\lambda_{\rm max} - \lambda_{\rm min}),$$

and contains N_v reciprocal-lattice points where $N_v = V_v/V^*$ and V^* is the volume of the reciprocal unit cell (Moffat, 1997). N_v can be large, particularly for crystals that diffract to high resolution and thus have larger values of $d^*_{\rm max}$. Laue patterns may therefore contain numerous closely spaced spots and exhibit a spatial-overlap problem (Cruickshank *et al.*, 1991). The value of N_v is up to an order of magnitude greater than the typical number of spots on a monochromatic oscillation pattern from the same crystal. Since the overall goal of a diffraction experiment is to record all spots in the unique volume of reciprocal space with suitable accuracy and redundancy, a Laue data set may contain fewer images and more spots of higher redundancy than a monochromatic data set (Clifton *et al.*, 1991). This is particularly evident if the crystal is of high symmetry.

Kalman (1979) provided derivations of the integrated intensity of a single spot in the Laue case and in the monochromatic case. Moffat (1997) used these to show that the duration of a typical Laue exposure was between three and four orders of magnitude less than the corresponding monochromatic exposure. The physical reason for this significant Laue advantage lies in the fact that all Laue spots are in a diffracting position and contribute to the integrated intensity throughout the exposure. In contrast, monochromatic spots diffract only briefly as each sweeps through the narrow Ewald sphere [more strictly, through the volume between the closely spaced Ewald spheres corresponding to $1/\lambda$ and $1/(\lambda + d\lambda)$]. The details are modified slightly for mosaic crystals of finite dimensions subjected to an X-ray beam of finite cross section and angular crossfire (Ren *et al.*, 1999; Z. Ren, unpublished results).

Exposure times are governed not merely by the requirement to generate sufficient diffracted intensity in a spot – the signal – but also to minimize the background under the spot – the noise. The background under a Laue spot tends to be higher than under a monochromatic spot, since it arises from a larger volume of reciprocal space in the Laue case. This volume extends from $d^*_{\rm min}$ (where $d^*_{\rm min} = 2\sin\theta/\lambda_{\rm max}$ and θ is the Bragg angle for that Laue spot) through the Laue spot at d^* to either $d^*_{\rm max}$ or $2\sin\theta/\lambda_{\rm min}$, whichever is the smaller (Moffat *et al.*, 1989). Since both the signal and the noise in a Laue pattern are directly proportional to the exposure time, their ratio is independent of that parameter. The ratio does depend on the wavelength range $(\lambda_{\rm max} - \lambda_{\rm min})$. Decreasing the wavelength range both generates fewer spots and increases the signal-to-noise ratio for each remaining spot by diminishing the background under it. This is analogous to decreasing the oscillation range in a monochromatic exposure.

The choice of appropriate exposure time in the Laue case is complicated, but the central fact remains: both in theory and in practice, Laue exposures are very short with respect to monochromatic exposures (Moffat *et al.*, 1984; Helliwell, 1985; Moffat, 1997). Satisfactory Laue diffraction patterns have been routinely obtained with X-ray exposures of 100 to 150 ps, corresponding to the duration of a single X-ray pulse emitted by a single 15 mA bunch of electrons circulating in the European Synchrotron Radiation Facility (ESRF) (Bourgeois *et al.*, 1996).

The advantages and disadvantages of the Laue technique, compared to the better-established and more familiar monochromatic techniques, are presented in Table 8.2.2.1.

8.2.3. Practical considerations in the Laue technique

The experimental aspects of a Laue experiment – the source and optics, the shutters and other beamline components, detectors, analysis software, and the successful design of the Laue experiments themselves – have been presented by Helliwell *et al.* (1989), Ren & Moffat (1994, 1995*a*,*b*), Bourgeois *et al.* (1996), Ren *et al.* (1996), Clifton *et al.* (1997), Moffat (1997), Yang *et al.* (1998), and Ren *et al.* (1999). Certain key parameters are under the experimenter's control, such as the nature of the source (bending magnet, wiggler or undulator), the wavelength range incident on the crystal (as modified by components of the beamline such as a mirror and attenuators), the choice of detector (active area, number of pixels and the size of each, dependence of detector parameters on wavelength, inherent background, and the accuracy and speed of readout), the experimental data-collection strategy (exposure time or times, number of angular settings of the crystal to be employed and the angular interval between them) and the data-reduction strategy (properties of the algorithms employed and of the software analysis package). A successful Laue experiment demands consideration of these parameters jointly and in advance, as described in these references. The goal is accurate structure amplitudes, not just speedily obtained, beautiful diffraction images.

For example, an undulator source yields a spectrum in which the incident intensity varies sharply with wavelength. Such a source should only be employed if the software can model this

Table 8.2.2.1

Advantages and disadvantages of the Laue technique

This table is adapted from Moffat (1997). See also Ren *et al.* (1999).

Advantages
Shortest possible exposure time, well suited to rapid time-resolved studies that require high time resolution.
Insensitive to all temporal fluctuations in the beam incident on the crystal, whether arising from the source itself, the optical components of the beamline or the shutter train. (Sensitive only to unusual fluctuations of the shape of the incident spectrum with time.)
All spots in a local region of the detector have an identical profile; none are (geometrically) partial.
Requires a stationary crystal and relatively simple optical components, therefore images are easy to acquire.
A large volume of reciprocal space is surveyed per image, hence fewer images are necessary to survey the entire unique volume.
High redundancy of measurements readily obtained, particularly at high resolution.
Disadvantages
Energy overlaps must be deconvoluted into their components if complete data are to be obtained, particularly at low resolution.
Spatial overlaps are numerous, particularly for mosaic crystals, and must be resolved.
Completeness at low resolution may be low, which would lead to significant series-termination errors in Fourier maps.
The rate of heating owing to X-ray absorption can be very high.
The wider the wavelength range, the higher the background under each spot; a trade-off is unavoidable between coverage of reciprocal space and accuracy of intensity measurements.
Spot shape is quite sensitive to crystal disorder.
More complicated wavelength-dependent corrections must be derived and applied to spot intensities to yield structure amplitudes.

variation suitably in the derivation of the wavelength normalization curve. This is indeed so (at least for the *LaueView* software package) even in the most extreme case, that of the so-called single-line undulator source in which $\lambda_{\max}$ and $\lambda_{\min}$ may differ by only 10%, say by 0.1 Å at 1.0 Å (V. Šrajer *et al.* and D. Bourgeois *et al.*, in preparation).

As a second example, a Laue diffraction pattern is particularly sensitive to crystal disorder, which leads to substantial 'streaking' of the Laue spots that is predominantly radial in direction in each diffraction image and may be dependent both on direction in reciprocal space (anisotropic disorder) and on time (if, in a time-resolved experiment, disorder is induced by the process of reaction initiation or by structural evolution as the reaction proceeds). The software therefore has to be able to model accurately elongated closely spaced or partially overlapping spots, whose profile varies markedly with position on each detector image and with time (Ren & Moffat, 1995*a*). If the software has difficulty with this task, then either a more ordered crystal must be selected, thus diminishing the size of each spot and the extent of spatial overlaps, or a narrower wavelength range must be used, thus reducing the total number of spots per image and their average spatial density (Cruickshank *et al.*, 1991); or the crystal-to-detector distance must be increased, thus increasing the average spot-to-spot distance and potentially increasing the signal-to-noise ratio. There are, however, trade-offs. More ordered crystals may not be readily available, a narrower wavelength range means that more images are required for a complete data set and the detector must continue to intercept all of the high-angle diffraction data (which consist largely of single spots stimulated by longer wavelengths).

As a third example, consider radiation damage. This can be purely thermal, arising from heating due to X-ray absorption. The rate of temperature rise may easily reach several hundred kelvin per second from a focused pink bending-magnet beam at second-generation sources such as the National Synchrotron Light Source (NSLS) (Chen, 1994; Moffat, 1997) or several thousand kelvin per second from a focused wiggler source at third-generation sources such as the ESRF. Fast shutters are required to provide an individual exposure of one millisecond or less in the latter case, and hence to limit the temperature rise to a readily survivable value of several kelvin (Bourgeois *et al.*, 1996; Moffat, 1997). Primary radiation damage (arising directly from X-ray absorption and hence from energy deposition) cannot be eliminated, but it may be modified by selection of the wavelength range and by lowering $\lambda_{\max}$. Secondary radiation damage, arising from the chemical and structural damage generated by highly reactive, rapidly diffusing free radicals, hydrated electrons and other chemical species, can be greatly minimized by the use of very short exposures which allow little time for damaging reactions to occur, and by working at cryogenic temperatures where diffusion is greatly reduced (see *e.g.* Garman & Schneider, 1997). However, the last strategy may not be an option in a time-resolved Laue experiment, where the desired structural transitions may be literally frozen out at cryogenic temperatures.

Extraction of structure amplitudes from a Laue image or data set proceeds through five stages, reviewed in detail by Clifton *et al.* (1997) and Ren *et al.* (1999), and outlined in Fig. 8.2.3.1. First comes the purely geometrical process of indexing, in which each spot is associated with the appropriate *hkl* value and the unit-cell parameters, crystal orientation matrix, $\lambda_{\min}$, geometric parameters of the detector and X-ray camera, $\lambda_{\max}$ and $d^*_{\max}$ are refined, also yielding λ, the wavelength stimulating that spot. In the second stage, each spot is integrated using appropriate profile-fitting algorithms. Thirdly, the wavelength normalization curve is derived, usually by comparison of the recorded intensities of the same (single) spots or symmetry-related spots at several crystal orientations, applied to each image, and the images in each data set scaled together. In the fourth stage, the intensities of spots identified in the first stage as multiple are resolved (or deconvoluted) into the intensity of each individual component or harmonic. The total intensity of a multiple Laue spot is the weighted sum of the intensities of each component, and the weights are known from the wavelength assigned to each component (Stage 1) and the wavelength normalization curve (Stage 3). In the fifth stage, the single and multiple data are merged and reduced to structure amplitudes. Ren *et al.* (1999) have emphasized that, in contrast to the simplified description presented in Section 8.2.2, the effective wavelength range $\lambda_{\max} - \lambda_{\min}$ depends on resolution, and each Laue spot is stimulated by a range of wavelengths which can be quite large at low resolution. Although typical Laue software packages such as the *Daresbury Laue Software Suite* (Helliwell *et al.*, 1989; Campbell, 1995), *LEAP* (Wakatsuki, 1993) and *LaueView* (Ren & Moffat, 1995*a*,*b*) are largely automated, a surprising degree of manual intervention may still be required in the first (indexing) stage, and in later stages where the order of parameter refine-

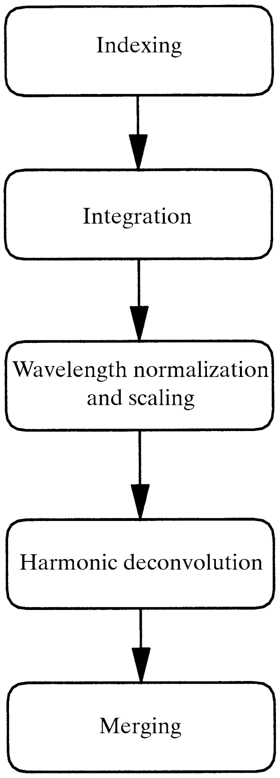


Figure 8.2.3.1
Flow chart of typical Laue data processing.

ment and various rejection criteria may be adjusted by the user. The overall result is that carefully conducted Laue experiments yield structure amplitudes that equal those from monochromatic data in quality (Ren *et al.*, 1999).

8.2.4. The time-resolved experiment

The principles and applications of time-resolved macromolecular crystallography have been widely reviewed (Moffat, 1989; Cruickshank *et al.*, 1992; Hajdu & Johnson, 1993; Helliwell & Rentzepis, 1997; Moffat, 1998; Ren *et al.*, 1999). This article therefore concentrates on the crystallographic aspects.

The essence of a perfect time-resolved crystallographic experiment is that a structural reaction is initiated in all the molecules in the crystal, rapidly, uniformly and in a non-damaging manner. The molecules, far from thermodynamic equilibrium immediately after the completion of the initiation process, relax through a series of structural transitions back to equilibrium. The course of the structural transitions is monitored through the time-dependence of the X-ray scattering. The structure amplitudes (and indeed the phases) associated with each Bragg peak *hkl* become time-dependent and may be denoted $|F(hkl, t)|$. The Fourier transform of the structure factors yields the time-dependent space-average structure of the molecules in the crystal. If all molecules behave independently of one another in the crystal, as they do in dilute solution, then the overall time dependence arises from the time dependence of the fractional populations of each time-independent structural state. That is, the crystal exhibits time-dependent substitutional disorder. (Lest this seem an unfamiliar concept, recall that a multi-site, partially occupied, heavy-atom derivative also exhibits substitutional disorder: the contents of each unit cell differ slightly, depending on whether a particular heavy-atom site is occupied in that unit cell or not. Such disorder fortunately does not invalidate the use of that derivative. The analysis proceeds as though a particular site were, say, 70% occupied in every unit cell of the crystal, although in reality that site is 100% occupied in 70% of the unit cells. In this example, the substitutional disorder is time-independent.) The crystal is thus imperfect: substitutional disorder breaks the translational symmetry. It follows that there must also be time-dependent non-Bragg scattering; but all studies to date have focused on the Bragg scattering.

The above describes a perfect experiment but, as might be expected, reality is different. Initiation techniques such as absorption of light from a laser pulse (Schlichting & Goody, 1997) unavoidably deposit energy in the crystal and give rise to a temperature jump and transient, reversible crystal disorder, evident as spot streaking. The magnitude of this temperature jump is proportional to the number of photons absorbed, which in turn is related to the concentration of photoactive species, the quantum yield for photoactivation and the fraction of molecules stimulated (Moffat, 1995, 1998). The necessity for limiting the magnitude of this temperature jump to retain crystallinity means that it is difficult to initiate the reaction in all molecules in the crystal. For example, photodissociation of carbon monoxide from carbonmonoxymyoglobin crystals was achieved in roughly 40% of the molecules (Šrajer *et al.*, 1996), and entry into the photocycle of photoactive yellow protein in roughly 20% of the molecules (Perman *et al.*, 1998). The magnitude of the time-dependent change in structure-factor amplitudes, given by

$$\Delta F(hkl, t) = |F(hkl, t)| - |F(hkl, 0)|,$$

is proportional to the fraction of molecules photoactivated and is therefore substantially diminished.

The main crystallographic challenge thus becomes the accurate determination of small values of $\Delta F(hkl, t)$ in the face of both random errors (arising from, for example, the small numbers of diffracted photons into the reflection *hkl* from a brief X-ray pulse) and systematic errors (arising from, for example, inaccurate determination of the Laue wavelength normalization curve, crystal-to-crystal scaling errors, inadequately corrected absorption effects, or time-dependent spot profiles). Precision is enhanced by acquiring highly redundant Laue data (mean redundancies typically between 5 and 15), which also afford an excellent measure of the variance of the structure amplitudes, and accuracy is enhanced by interleaving measurements of $|F(hkl, t)|$ with those of $|F(hkl, 0)|$ on the same crystal, at nearly the same time. Indeed, 'two-spot' Laue patterns may be acquired by recording both the $|F(hkl, t)|$ and $|F(hkl, 0)|$ diffraction patterns, slightly displaced with respect to each other, on the same detector [image plate or charge-coupled device (CCD)] prior to readout and quantification (Ren *et al.*, 1996). However, this doubles the background and halves the signal-to-noise ratio.

The values of $\Delta F(hkl, t)$ span a four-dimensional space. What is the best way to scan this four-dimensional space, having regard for the need to minimize errors? Interleaving measurements of $|F(hkl, t)|$ and $|F(hkl, 0)|$ has been achieved by fixing the delay time t between reaction initiation (the pump, laser pulse) and X-ray data acquisition (the probe, X-ray pulse), and surveying all values of *hkl* through progressive reorientation of the crystal between Laue images until all the unique volume of reciprocal space is surveyed with adequate redundancy and completeness (Šrajer *et al.*, 1996; Perman *et al.*, 1998; Ren *et al.*, 1999). The value of t is then altered and data acquisition repeated for all suitable values of t. That is, t is the slow variable. The difficulty with this approach is that a single crystal may yield only one or two data sets, corresponding to one or two values of t, before radiation damage (predominantly laser-induced rather than X-ray-induced) compels replacement of the crystal. The entire reaction time course over all values of t must therefore be pieced together from measurements on many crystals, a process which is prone to inter-crystal scaling errors. A second experimental approach to scanning this four-dimensional data space is therefore to fix the crystal orientation, obtain values of $|F(hkl, t)|$ and $|F(hkl, 0)|$ for all suitable values of t, reorient the crystal, recollect these same values of t, then replace the crystal and repeat until all the unique volume of reciprocal space is surveyed. That is, *hkl* are the slow variables (B. Perman, S. Anderson & Z. Ren, unpublished results). This approach yields a more accurate time course, but (for a single crystal) from a subset of reflections only. In practice, of the order of 100 time points t may be collected.

Table 8.2.5.1
Time-resolved Laue diffraction experiments

This table is adapted from Table 2 of Ren *et al.* (1999), in which citations of the original experiments are provided.

Protein	Time resolution	Experiment
Hen lysozyme	64 ms	Temperature jump test
Glycogen phosphorylase	1 s	Bound maltoheptose
Hen lysozyme	1 s	Radiation damage test
Glycogen phosphorylase	100 ms	Use of caged phosphate
Ras oncogene product	1 s	GTP complex
γ-Chymotrypsin	5 s	Photolysis of cinnamate/pyrone
Trypsin	800 ms	Ordered hydrolytic water
Cytochrome *c* peroxidase	1 s	Redox active compound I
Hen lysozyme	10 ms	Temperature jump
Isocitrate dehydrogenase	50 ms	ES complex and intermediate
Isocitrate dehydrogenase	10 ms	Product complex
Photoactive yellow protein	10 ms	*p*B-like intermediate
Photoactive yellow protein	10 ns	*p*R-like intermediate
CO-myoglobin	10 ns	Photolyzed CO species at 290 K
CO-myoglobin	8 ms	Photolyzed CO species at 20–40 K
Hydroxymethylbilane synthase	1.5 ms	Mutant enzyme–cofactor complex

The first approach permits Fourier or difference Fourier maps to be calculated using data from a single crystal at one or a small number of time delays t. The second approach requires data from many crystals to be acquired before such maps can be calculated. This complicates the issue of which is the better approach. Preliminary results (V. Šrajer & B. Perman, unpublished results) suggest that genuine features may be reliably distinguished in real space by examination of Fourier and difference Fourier maps, but genuine trends in reciprocal space are much harder to discern.

How many features can be distinguished as genuine in real space, in, for example, a difference Fourier map? We presently employ three criteria. First, the feature must be 'significant' in crystallographic terms. That is, its peak height must exceed (say) 4σ to 5σ, where σ is the r.m.s. value of the difference electron density $\Delta\rho$ across the asymmetric unit, in a difference Fourier map. Second, the feature must be chemically plausible, *e.g.* located on or near critical groups in the active site. Third, the feature must persist over several time points. No genuine feature is likely to vary faster than exponentially in time (though slower variation is possible), but noise features tend to come and go, varying rapidly with time. The third criterion, which in effect is applying a low-pass temporal filter to the data, or 'time-smoothing', is only applicable if several time points are available per decade of time t. It is to ensure that this powerful criterion can be applied in an unbiased manner that the time points t at which data are acquired are uniformly and closely spaced in log t.

Suppose that complete and accurate values of $|F(hkl, t)|$ are available to high resolution and at numerous values of t. How can these time-dependent data be further analysed to yield information on the reaction mechanism and the time-dependent structures of intermediates? Each candidate chemical-kinetic mechanism implies a different time-dependent mixture of structural states at all times t. For each mechanism, a set of trial time-dependent intermediate structures can be calculated from the time-dependent data (Perman, 1999). One then asks: Is each trial intermediate structure an authentic, single, stereochemically plausible, refinable protein structure? If so, the mechanism is supported, but if not, the mechanism is rejected. This process, of seeking to extract time-independent structures from time-dependent data, is closely related to the better-understood process of extracting time-independent difference spectra from time-dependent optical absorption data *via*, for example, singular value decomposition or principal component analysis. The latter, optical analysis, proceeds in two dimensions, $\mathrm{OD}(\lambda, t)$; the former, crystallographic analysis, must proceed in four dimensions, either $\rho(xyz, t)$ or $|F(hkl, t)|$.

It will be appreciated that the acquisition of fast, time-resolved data is greatly hindered by the lack of a time-slicing area detector. This lack is even more evident when the structural reaction is irreversible as, for example, in the photoactivation of caged GTP to GTP (Schlichting *et al.*, 1990). In such cases, the reactants must be replenished prior to each reaction initiation, which makes the acquisition of time-resolved data particularly tedious. The present generation of CCD detectors have an inter-frame time delay in the millisecond (or just sub-millisecond) time range. Pixel array detectors under development may permit the acquisition of sequential images with a time delay in the microsecond range. The desirable nanosecond or even picosecond time range seems inaccessible for area detectors (but not for point detectors such as streak cameras). A new approach may be needed, such as the use of chirped hard X-ray pulses which, in combination with Laue diffraction, map X-ray energy into both reciprocal space (*hkl*) and time (K. Moffat, in preparation).

8.2.5. Conclusions

Only a small number of biochemical systems have been subjected to time-resolved crystallographic analysis (Table 8.2.5.1; Ren *et al.*, 1999). The experiments are technically demanding, require careful planning in the execution, in data analysis and in data interpretation, and strategies for the evaluation of mechanism are still being developed. However, road maps exist for several successful classes of experiments (see *e.g.* Stoddard *et al.*, 1998; Moffat, 1998; Ren *et al.*, 1999) and new biological systems to which such analyses may be readily applied are being developed. In a world of structural genomics where structures themselves are ten-a-penny, a structure-based understanding of mechanism at the chemical level is still rare. The contributions of crystallography to functional – not merely structural – genomics may be large indeed.

This work was supported by the NIH. I thank Zhong Ren for comments on the manuscript.

References

Amorós, J. L., Buerger, M. J. & Canut de Amorós, M. (1975). *The Laue Method.* New York: Academic Press.

Bourgeois, D., Ursby, T., Wulff, M., Pradervand, C., Legrand, A., Schildkamp, W., Labouré, S., Srajer, V., Teng, T. Y., Roth, M. & Moffat, K. (1996). *Feasibility and realization of single-pulse Laue diffraction on macromolecular crystals at ESRF. J. Synchrotron Rad.* **3**, 65–74.

Campbell, J. W. (1995). *LAUEGEN, an X-windows-based program for the processing of Laue X-ray diffraction data. J. Appl. Cryst.* **28**, 228–236.

Cassetta, A., Deacon, A., Emmerich, C., Habash, J., Helliwell, J. R., McSweeney, S., Snell, E., Thompson, A. W. & Weisgerber, S. (1993). *The emergence of the synchrotron Laue method for rapid data collection from protein crystals*. *Proc. R. Soc. London Ser. A*, **442**, 177–192.

Chen, Y. (1994). PhD thesis, Cornell University, USA.

Clifton, I. J., Duke, E. M. H., Wakatsuki, S. & Ren, Z. (1997). *Methods Enzymol.* **277**, 448–467.

Clifton, I. J., Elder, M. & Hajdu, J. (1991). *Experimental strategies in Laue crystallography. J. Appl. Cryst.* **24**, 267–277.

Cruickshank, D. W. J., Helliwell, J. R. & Johnson, L. N. (1992). *Time-Resolved Macromolecular Crystallography*. Oxford: Oxford Science Publications.

Cruickshank, D. W. J., Helliwell, J. R. & Moffat, K. (1987). *Multiplicity distribution of reflections in Laue diffraction*. *Acta Cryst.* A**43**, 656–674.

Cruickshank, D. W. J., Helliwell, J. R. & Moffat, K. (1991). *Angular distribution of reflections in Laue diffraction. Acta Cryst.* A**47**, 352–373.

Garman, E. F. & Schneider, T. R. (1997). *Macromolecular cryocrystallography. J. Appl. Cryst.* **30**, 211–237.

Hajdu, J. & Johnson, L. N. (1993). *Biochemistry*, **29**, 1669–1675.

Helliwell, J. R. (1984). *Synchrotron X-radiation protein crystallography: instrumentation, methods and applications*. *Rep. Prog. Phys.* **47**, 1403–1497.

Helliwell, J. R. (1985). *Protein crystallography with synchrotron radiation*. *J. Mol. Struct.* **130**, 63–91.

Helliwell, J. R., Habash, J., Cruickshank, D. W. J., Harding, M. M., Greenhough, T. J., Campbell, J. W., Clifton, I. J., Elder, M., Machin, P. A., Papiz, M. Z. & Zurek, S. (1989). *The recording and analysis of synchrotron X-radiation Laue diffraction photographs. J. Appl. Cryst.* **22**, 483–497.

Helliwell, J. R. & Rentzepis, P. M. (1997). *Time Resolved Diffraction.* Oxford University Press.

Kalman, Z. H. (1979). *On the derivation of integrated reflected energy formulae. Acta Cryst.* A**35**, 634–641.

Moffat, K. (1989). *Time-resolved macromolecular crystallography*. *Annu. Rev. Biophys. Biophys. Chem.* **18**, 309–332.

Moffat, K. (1995). *Proc. Soc. Photo-Opt. Instrum. Eng.* **2521**, 182–187.

Moffat, K. (1997). *Laue diffraction*. *Methods Enzymol.* **277B**, 433–447.

Moffat, K. (1998). *Ultrafast time-resolved crystallography*. *Nature Struct. Biol.* **5**, 641–643.

Moffat, K., Bilderback, D., Schildkamp, W., Szebenyi, D. & Teng, T.-Y. (1989). In *Synchrotron Radiation in Structural Biology*, edited by R. M. Sweet and A. J. Woodhead, pp. 325–330. New York and London: Plenum Press.

Moffat, K., Szebenyi, D. & Bilderback, D. (1984). *X-ray Laue diffraction from protein crystals*. *Science*, **223**, 1423–1425.

Perman, B. (1999). PhD thesis, University of Chicago, USA.

Perman, B., Šrajer, V., Ren, Z., Teng, T.-Y., Pradervand, C., Ursby, T., Bourgeois, D., Schotte, F., Wulff, M., Kort, R., Hellingwerf, K. & Moffat, K. (1998). *Energy transduction on the nanosecond time scale: early structural events in a xanthopsin photocycle*. *Science*, **279**, 1946–1950.

Ren, Z., Bourgeois, D., Helliwell, J. R., Moffat, K., Šrajer, V. & Stoddard, B. L. (1999). *Laue crystallography: coming of age. J. Synchrotron Rad.* **6**, 891–917.

Ren, Z. & Moffat, K. (1994). *Laue crystallography for studying rapid reactions. J. Synchrotron Rad.* **1**, 78–82.

Ren, Z. & Moffat, K. (1995*a*). *Quantitative analysis of synchrotron Laue diffraction patterns in macromolecular crystallography. J. Appl. Cryst.* **28**, 461–481.

Ren, Z. & Moffat, K. (1995*b*). *Deconvolution of energy overlaps in Laue diffraction. J. Appl. Cryst.* **28**, 482–493.

Ren, Z., Ng, K., Borgstahl, G. E. O., Getzoff, E. D. & Moffat, K. (1996). *Quantitative analysis of time-resolved Laue diffraction patterns. J. Appl. Cryst.* **29**, 246–260.

Schlichting, I., Almo, S. C., Rapp, G., Wilson, K., Petratos, K., Lentfer, A., Wittinghofer, A., Kabsch, W., Pai, E. F., Petsko, G. A. & Goody, R. S. (1990). *Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis*. *Nature (London)*, **345**, 309–315.

Schlichting, I. & Goody, R. S. (1997). *Methods Enzymol.* **277B**, 467–490.

Šrajer, V., Teng, T.-Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Schildkamp, W., Bourgeois, D., Wulff, M. & Moffat, K. (1996). *Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography*. *Science*, **274**, 1726–1729.

Stoddard, B. L., Cohen, B. E., Brubaker, M., Mesecar, A. D. & Koshland, D. E. Jr (1998). *Millisecond Laue structures of an enzyme-product complex using photocaged substrate analogues*. *Nature Struct. Biol.* **5**, 891–897.

Wakatsuki, S. (1993). In *Data Collection and Processing*, edited by L. Sawyer, N. W. Isaacs & S. Bailey, pp. 71–79. DL/Sci/R34. Warrington: Daresbury Laboratory.

Wood, I. G., Thompson, P. & Mathewman, J. C. (1983). *A crystal structure refinement from Laue photographs taken with synchrotron radiation. Acta Cryst.* B**39**, 543–547.

Yang, X., Ren, Z. & Moffat, K. (1998). *Structure refinement against synchrotron Laue data: strategies for data collection and reduction. Acta Cryst.* D**54**, 367–377.

International Tables for Crystallography (2012). Vol. F, Chapter 9.1, pp. 211–230.

PART 9. X-RAY DATA COLLECTION

Chapter 9.1. Principles of monochromatic data collection

Z. Dauter and K. S. Wilson

9.1.1. Introduction

X-ray data collection is the central experiment in a crystal structure analysis. For small-molecule structures, the availability of intensity data to atomic, usually around 0.8 Å, resolution means that the phase problem can be solved directly and the atomic positions refined with a full anisotropic model. This results in a truly automatic structure solution for most small molecules.

Macromolecular crystals pose much greater problems with regard to data collection. The first arise from the size of the unit cell, resulting in lower average intensities of individual reflections coupled with a much greater number of reflections (Table 9.1.1.1). Secondly, the crystals usually contain considerable proportions of disordered aqueous solvent, giving further reduction in intensity at high resolution and, in the majority of cases, restricting the resolution to be much less than atomic. Thirdly, the crystals are sensitive to radiation damage (see Section 9.1.12). Such problems have severe implications for all subsequent steps in a structure analysis. Solution of the phase problem is generally not possible through direct methods, except for a small number of exceptionally well diffracting proteins. The refined models require the imposition of stereochemical constraints or restraints to maintain an acceptable geometry.

At modern synchrotron beamlines, cryogenic cooling and high-efficiency two-dimensional (2D) detectors have made data collection technically easier, but it remains a fundamental scientific procedure underpinning the whole structural analysis. Therefore, it is essential to take the greatest care over this key step. The aim of this chapter is to indicate procedures for optimizing data acquisition. Overviews on several issues related to this topic have been published (Carter & Sweet, 1997; Evans & Walsh, 2005; Dauter, 2005).

9.1.2. The components of a monochromatic X-ray experiment

To collect X-ray data from single crystals, the following elements are required:

(1) a source of X-rays;
(2) optical elements to focus the X-rays onto the sample;
(3) a monochromator to select a single wavelength;
(4) a collimator to produce a beam of defined dimension;
(5) a shutter to limit the exposure of the sample to X-rays;
(6) a goniostat with associated sample holder to allow rotation of the crystal;
(7) the crystalline sample itself;
(8) a cryogenic cooling device for vitrified crystals;
(9) an efficient, generally 2D, detector system;
(10) software to control the experiment and store and display the X-ray images;
(11) data-processing software to extract intensities and associated standard uncertainties for the Bragg reflections in the images.

On a number of beamlines, automated procedures have been implemented for sample changing, automatic crystal centring, evaluating the diffraction and proposing a strategy for data collection. These allow rapid and more effective selection of the best sample and optimal parameters.

Many of these are discussed elsewhere in this volume. This chapter aims to provide guidance in those areas where choices are to be made by the experimenter and is concerned with the interrelations between parameters and how they conspire for or against different strategies of data collection.

9.1.3. Data completeness

The advantage of diffraction methods over spectroscopy is that they provide a full 3D view of the object. Diffraction methods are theoretically limited by the wavelength of the radiation used, but, in practice, every diffraction experiment is further limited by the aperture and quality of the lens. In the X-ray experiment, the aperture corresponds to the resolution limit and the quality of the 'lens' to the completeness and accuracy of the measured Bragg reflection intensities.

In this context, completeness has two components, the first of which is geometric and hence quantitative. It is necessary to rotate the crystal so that all unique reciprocal-lattice points pass through the Ewald sphere and the associated intensities are recorded on the detector. Ideally, the intensities of 100% of the unique Bragg reflections should be measured. The second component is qualitative and statistical: for each *hkl*, the intensity, I_{hkl}, should be significant, with its accuracy correctly estimated in the form of an associated standard uncertainty, $\sigma(I)$. The data should be significant in terms of the $I/\sigma(I)$ ratio throughout the resolution range. This point will be returned to below, but it is especially important that the data at low resolution are complete and not overloaded on the detector, and that there is not a significant fraction of essentially zero-level intensities in the higher-resolution shells.

9.1.4. X-ray sources

There are two principal sources of X-rays appropriate for macromolecular data collection: rotating anodes and synchrotron storage rings. These are not discussed in detail here as they are described in Chapters 6.1 and 8.1.

Table 9.1.1.1
Size of the unit cell and number of reflections

Compound	Unit cell Edge (Å)	Volume (Å^3)	Reflections	Average intensity
Small organic	10	1000	2000	1
Supramolecule	30	25000	30000	1/25000
Protein	100	1000000	100000	1/1000000
Virus	400	100000000	1000000	1/100000000

Copyright © 2012 International Union of Crystallography

9.1.4.1. *Conventional sources*

Rotating anodes were initially developed for biological scattering experiments on muscle samples and have the advantage of higher intensity compared with sealed-tube generators. They usually have a copper target providing radiation at a fixed wavelength of 1.542 Å. Alternative targets, such as silver or molybdenum, provide lower intensities at short wavelengths, but have not found general applications to macromolecules. It is also possible to use a chromium target, giving a longer wavelength of 2.29 Å. Historically, rotating anodes were first used with nickel filters to give monochromatic Cu *Kα* radiation. Current systems are equipped with either graphite monochromators, a focusing mirror or multilayer optics. The latter provide substantially enhanced intensity. Rotating anodes remain the source of choice in most structural biology laboratories. An important choice for the user is in the selection of optimal collimator aperture: this should roughly match the crystal sample dimensions. For large crystals, especially if the cell dimensions are also large, it may be preferable to use collimator settings smaller than the crystal in order to resolve the diffraction spots on the detector. Considerable progress has been made in the technologies for rotating anodes in recent years, as well as in the development of high-intensity sealed-tube sources for home laboratories (Bloomer & Arndt, 1999).

9.1.4.2. *Synchrotron storage rings*

The radiation intensity available from rotating anodes is limited by the heat load per unit area on the target. In the early 1970s, it was realized that synchrotron storage rings produced X-radiation in the necessary spectral range for studies in structural molecular biology (Rosenbaum *et al.*, 1971), and the last four decades have seen great advances in their application to macromolecular crystallography (Helliwell, 2004; Hendrickson, 2000). Synchrotron radiation (SR) is now used for the great majority of newly determined protein-crystal structures.

The general advantages of SR are:

(1) High intensity: third-generation sources provide more than 1000 times the intensity of a conventional source.
(2) A highly parallel beam allowing the resolution of closely spaced spots from large unit cells.
(3) Short wavelengths, less than 1 Å, essentially eliminating the problems of correcting for absorption.
(4) Tunability of the wavelength, allowing its optimization for single- or multiple-wavelength applications; this is simply not possible with a conventional source.
(5) The ability to use a white, non-monochromated beam, the so-called Laue technique discussed in Chapter 8.2.

SR beamlines take a number of forms. The source may be a bending magnet or an insertion device, such as a wiggler or an undulator. The properties of different beamlines thus vary considerably and it is vital to choose an appropriate beamline for any particular application. The beamline capabilities are, of course, affected by the detector as well as the source itself. As far as the user is concerned, the primary questions regard the intensity, the size of the focal spot, the wavelength tunability and the detector system.

The present consensus for new synchrotron beamlines for macromolecular crystallography is that they should be on sources with an energy of at least 3 GeV and should receive radiation from tunable undulators. Together, these provide high and tunable intensity over the range required for most crystallographic experiments, including multiwavelength anomalous dispersion (MAD). The impact of free-electron lasers, which are currently under construction at a number of sites, is not yet possible to assess.

Present beamlines produce radiation of extremely high quality for macromolecular data collection. At third-generation sources complete data sets can be collected from cryogenically vitrified single crystals in minutes.

9.1.5. Goniostat geometry

9.1.5.1. *Overview*

The diffraction condition for a particular reflection is fulfilled when the corresponding reciprocal-lattice point lies on the surface of the Ewald sphere. If a stationary crystal is irradiated by the X-ray beam, only a few reflections will lie in the diffracting position. To record intensities of a larger number of reflections, either the size of the Ewald sphere or the crystal orientation has to be changed. The first option, with the use of non-monochromatic, or 'white', radiation, is the basis of the Laue method (Chapter 8.2). If the radiation is monochromatic, *i.e.* single-wavelength, the crystal has to be rotated during exposure to bring successive reflections into the diffraction condition.

9.1.5.2. *The screenless rotation method and 2D detectors*

In the early days of protein crystallography, a number of geometries were used for X-ray cameras, notably the Weissenberg and precession methods. In addition, single-counter diffractometers were used with four-circle goniostats. However, in practice only the screenless rotation geometry (Arndt & Wonacott, 1977) survives today. This requires a 2D detector, which was initially in the form of photographic film. It is of significance that typical film sizes were of the order of 10 × 10 cm with up to 2000 × 2000 scanned pixels; a similar effective area has proved effective for image plates and charge-coupled devices (CCDs).

However, automation of protein-data collection needed efficient 2D detectors (Part 7). The first were multiwire proportional counters, which found widespread use in the early 1980s (Hamlin, 1985). These finally proved to be limited by a combination of spatial resolution and dead time of the read-out. A major advance occurred in the late 1980s with the widespread introduction of imaging plates (Amemiya, 1995), scanned on-line both at synchrotron beamlines and on laboratory rotating-anode sources. This represented a revolution in macromolecular data collection, making it technically straightforward to save full 2D images with sufficient positional resolution and dynamic range to computer disk automatically. The limiting factor of the imaging plate proved to be the slow read-out time of the order of several seconds to minutes. At high-intensity sources in particular, *e.g.* third-generation SR sites, exposure times per image can fall to one second or less, and with an imaging plate the bulk of the time is spent reading the detector image rather than collecting data. Typical data-collection times with imaging plates remained of the order of several hours, even with the use of SR. This is a much smaller problem with rotating-anode sources, where exposure times dominate the duty cycle.

For high-intensity SR sites, the detector of choice is the CCD (Gruner & Ealick, 1995). The spatial resolution is comparable with that of imaging plates, but the read-out time can be as low as one to two seconds. This means that complete data can be

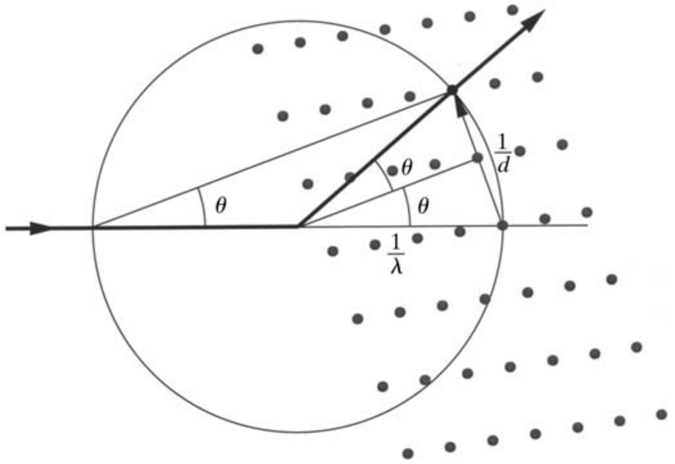


Figure 9.1.6.1
The Ewald-sphere construction. A reciprocal-lattice point lies on the surface of the sphere if the following trigonometric condition is fulfilled: $1/2d = (1/\lambda)\sin\theta$. After a simple rearrangement, it takes the form of Bragg's law: $\lambda = 2d\sin\theta$. Therefore, when a reciprocal-lattice point with indices *hkl* lies on the surface of the Ewald sphere, the interference condition for that particular reflection is fulfilled and it gives rise to a diffracted beam directed along the line joining the centre of the sphere to the reciprocal-lattice point on the surface.

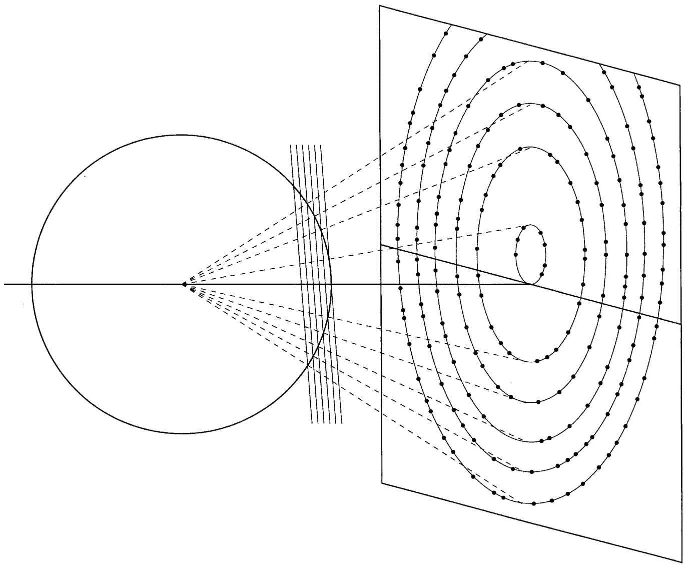


Figure 9.1.6.2
The plane of reflections in the reciprocal sphere that is approximately perpendicular to the X-ray beam gives rise to an ellipse of reflections on the detector.

recorded in minutes rather than hours and has transformed approaches to data collection.

While the CCD has revolutionized data-collection times, further advances are expected from the use of solid-state pixel detectors. Such detectors record individual X-ray quanta and have essentially zero read-out time. The most advanced of these, the PILATUS 1M device, is a hybrid pixel array detector (Broennimann *et al.*, 2006; Hülsen *et al.*, 2006), first installed at the Swiss Light Source.

Almost all current 2D detectors are used in conjunction with a goniostat, providing rotation of the crystal about a single axis during exposure. Indeed, the majority of instruments have only a single rotation axis. The remainder are based on the kappa (ω, κ, φ) cradle to select different initial orientations of the sample in the beam; the sample is nevertheless subsequently rotated about a single axis for data collection.

9.1.6. Basis of the rotation method

9.1.6.1. Rotation geometry

The physical process of diffraction from a crystal involves the interference of X-rays scattered from the electron clouds around the atomic centres. The ordered repetition of atomic positions in all unit cells leads to discrete peaks in the diffraction pattern. The geometry of this process can alternatively be described as resulting from the reflection of X-rays from a set of hypothetical planes in the crystal. This is explained by the Ewald construction (Fig. 9.1.6.1), which provides a visualization of Bragg's law. Monochromatic radiation is represented by a sphere of radius $1/\lambda$, and the crystal by a reciprocal lattice. The lattice consists of points lying at the end of vectors normal to reflecting planes, with a length inversely proportional to the interplanar spacing, $1/d$. In the rotation method, the crystal is rotated about a single axis, with the rotation angle defined as φ. A seminal work, giving an excellent background to this field by a number of contributors, was edited by Arndt & Wonacott (1977). A more recent review of the method is provided by Dauter (2005).

9.1.6.2. Diffraction pattern at a single orientation: the 'still' image

For a stationary crystal in any particular orientation (a so-called 'still' exposure), only a fraction of the total number of Bragg reflections will satisfy the diffracting condition. The number of reflections will be very limited for a small-molecule crystal, possibly zero in some orientations. Macromolecules have large unit cells, of the order of 100 Å, compared with the wavelength of the radiation, which is about 1.0 Å. In geometric terms, the reciprocal space is densely populated with points in relation to the size of the Ewald sphere. Thus, more reflections diffract simultaneously but at different angles, since many reciprocal-lattice points (reflections) lie simultaneously on the surface of the Ewald sphere in any crystal orientation. This is the great advantage of 2D detectors for large cell dimensions.

The real crystal is a regular and ordered array of unit cells. This means that reciprocal space is made up of a set of points organized in regular planes. For a still exposure, any particular plane of points in the reciprocal lattice intersects the surface of the Ewald sphere in the form of a circle. The corresponding diffracted rays, originating from the centre of the Ewald sphere, form a cone that intersects the sphere on the circle formed by the set of points. In most experiments, the detector is placed perpendicular to the direct beam and the cone of diffracted rays forms an ellipse of spots on its surface (Fig. 9.1.6.2). If a major axis of the crystal lies nearly parallel to the beam, then the ellipses will approximate a set of circles around the centre of the detector. All reflections within each circle will have one index in common, corresponding to the unit-cell axis lying along the beam. For non-centred unit cells, this index will increase by one in successive circles. The gaps between the circles depend on the spacing between the members of the set of reciprocal-lattice planes and are inversely proportional to the real cell dimension related to these planes.

Still exposures were used extensively in the early applications of the rotation method for estimation of crystal alignment. The geometric location of the spots with respect to the origin allows accurate determination of the unit-cell parameters and the crystal orientation. This approach has been superseded in modern

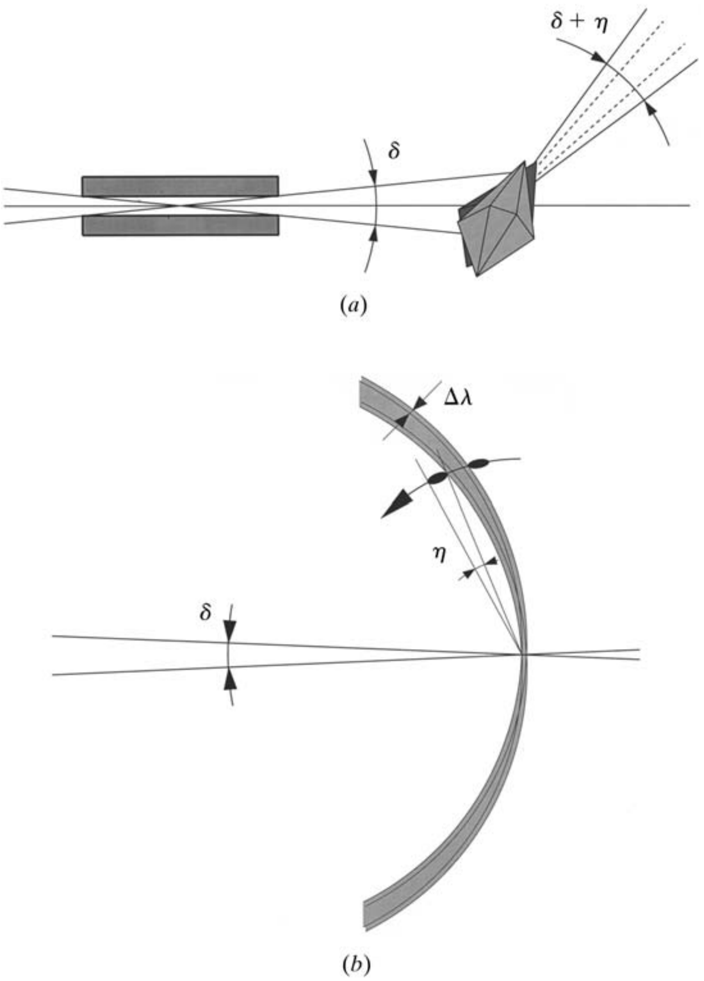


Figure 9.1.6.3
Schematic representation of beam divergence (δ) and crystal mosaicity (η). (*a*) In direct space, (*b*) in reciprocal space, where the additional thickness of the Ewald sphere results from the finite wavelength bandpass, $\delta\lambda/\lambda$.

software packages by autoindexing algorithms using real rotation images.

9.1.6.3. Rocking curve: crystal mosaicity and beam divergence

The Ewald-sphere construction assumes an ideal source with a totally parallel X-ray beam and an ideal crystal with all unit cells having identical relative orientation, resulting in infinitely sharp Bragg reflections. These assumptions lead to a sphere of radius $1/\lambda$ attached rigidly to the beam and with the crystal in a particular orientation as a reciprocal lattice consisting of mathematical points. A real experiment deviates from this in three respects. Firstly, the incident beam is not strictly parallel. On a conventional rotating-anode source the beam can only be focused and collimated to be parallel within a small angle, with a divergence of about 0.2° (with mirror optics) and 0.4° (with a monochromator). On SR sources, a much smaller beam divergence can be achieved, and, indeed, beamlines on third-generation SR sources approach the ideal ever more closely. The horizontal and vertical beam divergence may differ, and this must be taken into account. The Ewald sphere now has two limiting orientations which result in a nonzero active width. Secondly, the X-radiation is only monochromatic within a defined wavelength bandpass, $\delta\lambda/\lambda$, of the order 0.0002–0.001 at synchrotron beamlines, but considerably more for laboratory sources. The wavelength bandpass, in effect, thickens the surface of the Ewald sphere. Thirdly, real crystals are made up of small mosaic blocks imperfectly oriented relative to one another, increasing the total rocking curve. At room temperature, protein crystals often show a mosaic spread less than 0.05°, but for some samples this may be much larger. However, vitrification of crystals in many cases leads to substantial increase of mosaicity to sometimes more than 1°. In the reciprocal lattice, the effect of this is to give a finite dimension to each of the lattice points.

These effects are schematically illustrated in Fig. 9.1.6.3. The combined result is that the diffraction of a particular reflection is spread over a range of crystal rotation.

9.1.6.4. Rotation images and lunes

Using monochromatic radiation, in order to measure the remaining reflections that do not lie on the surface of the sphere, the crystal must be rotated to bring the reflections into the diffracting condition. If the crystal is rotated about a single axis during sequential exposures, this is known as the rotation method. The rotation axis is, in practice, chosen to be perpendicular to the beam to preserve the symmetry between the two halves of the complete pattern. This is the most commonly applied method of data collection for macromolecular crystals (Arndt & Wonacott, 1977).

If the crystal is rotated during exposure, the ellipses observed on a still image change their position on the detector. In effect, all reflections diffracting during one exposure will be contained within lunes formed between the two limiting positions of each ellipse at the start and end of the given rotation. The width of the lunes in the direction of the crystal rotation, perpendicular to the rotation axis, is proportional to the rotation range per exposure. In contrast, along the rotation axis the width of the lunes is very small, since the intersection of the reciprocal-lattice plane with the Ewald sphere does not change significantly. For crystals of small molecules, the lunes are not pronounced, owing to the sparse population of reciprocal space, but for crystals with large cell dimensions, the lunes are densely populated by diffraction spots and often exhibit clear and well pronounced edges. At high resolution, the mapping of the reciprocal lattice within each lune is distorted, and rows of reflections form hyperbolas. At low diffraction angles, where the surface of the Ewald sphere is approximately flat, this distortion is minimal, and the lunes look like fragments of precession photographs.

9.1.6.5. Partially and fully recorded reflections

The rotation method gives rise to lunes of data between the ellipses that relate to the start and the end of the rotation range used for the exposure. The data are complete if the Ewald sphere has been crossed by all reflections in the asymmetric part of the reciprocal lattice, which means that the crystal has to be rotated in total by a substantial angle. However, it is impossible to record all the data in a single exposure with such a wide rotation, owing to overlapping of the diffraction spots.

In practical applications to macromolecules, the total rotation is divided into a series of narrow individual rotations of width $\Delta\varphi$. In each of these, the crystal is exposed for a specified time or X-ray dose per angular unit. Each reflection diffracts over a defined crystal rotation and hence time interval, owing to the finite value of the rocking curve or angular spread, here referred to as ξ, the combined effect of beam divergence (δ) and crystal mosaicity (η). Provided ξ is less than $\Delta\varphi$, some reflections will start and finish crossing the Ewald sphere and hence diffract within one exposure. Their full intensity will be recorded on a

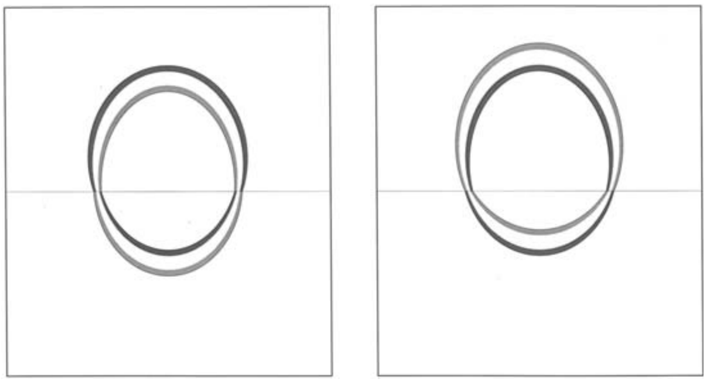
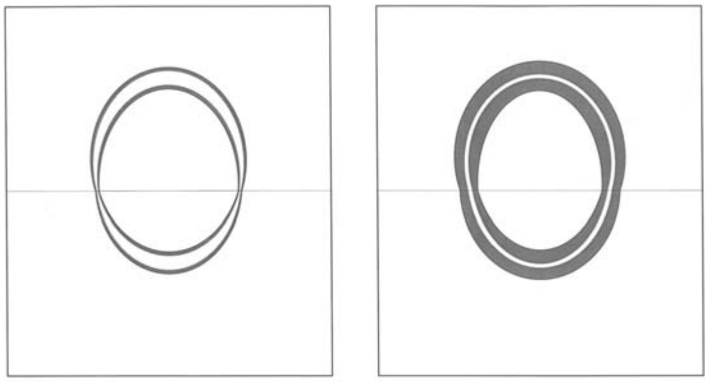


Figure 9.1.6.4
A single lune on two consecutive exposures. The partial reflections appear on both images and their intensity is distributed over both.



$\eta < \Delta\varphi$ $\eta \sim \Delta\varphi$

Figure 9.1.6.5
Appearance of a lune for (left) a crystal of low mosaicity and (right) a highly mosaic crystal. Characteristically, the width of the lune along the rotation axis is wider if the mosaicity is high.

single image, and these are referred to as fully recorded reflections, or fullys.

Other reflections will start to diffract during one exposure, but will still be diffracting at the end of the $\Delta\varphi$ rotation range. The remaining intensity of these reflections will be recorded on subsequent images. There will of course be corresponding reflections at the start of the present image. These reflections are termed partially recorded, or partials. Fig. 9.1.6.4 shows schematically how a lune appears on two consecutive exposures, with partials at each edge. The partials at the bottom edge of each lune contain the rest of the intensity of the partials from the previous exposure. The rest of the intensity of the partials at the top of the lune will appear on the next exposure. Superposition of two successive images will reveal some spots common to both: they are the partials shared between the two. If the angular spread ξ is small compared with the rotation range $\Delta\varphi$ then most reflections will be fully recorded. As ξ increases, the proportion of partials will rise, and when it reaches or exceeds $\Delta\varphi$ in magnitude there will be no fully recorded reflections. If the rotation range per image is small compared with the rocking curve, individual reflections can be spread over several images.

As ξ increases, the lunes become wider (Fig. 9.1.6.5), since there are more partial reflections crossing the Ewald sphere at any one time. The appearance of the lunes can be used to estimate the mosaicity of the crystal. If the edges are sharply defined, then the mosaicity is low. In contrast, if the intensities at the edges gradually fade away, then the mosaicity must be high. Indeed, this phenomenon can be exploited by the integration software to provide accurate definition of the orientation parameters and of the mosaicity.

A key characteristic of high mosaicity is that all lunes are wide in the region along the rotation axis. On still exposures, the width of the rings is proportional to the angular spread. The width of lunes is expected to be very small along the rotation axis. If they are wide in this region, this is especially indicative of high mosaic spread. While highly ordered crystals with low mosaicity are preferable and often lead to data of the highest quality, high mosaic spread is not a prohibitive factor in accurate intensity estimation, provided it is properly taken into account in estimating the data collection and integration parameters, such as individual rotation ranges.

9.1.6.6. The width of the rotation range per image: fine φ slicing

An important variable in the rotation method is the width of the rotation ranges per individual exposure. The two basic approaches can be termed wide and fine φ slicing, and differ in the relation between the angular spread and the rotation range per exposure. The two methods are applicable under different experimental constraints.

Fine φ slicing requires that the individual intensities are divided over several consecutive images, *i.e.* $\Delta\varphi$ should be substantially less than ξ (Kabsch, 1988). This approach possesses two very positive features. Firstly, it minimizes the background by integrating intensities only over a φ range equivalent to the rocking curve of the crystal. Secondly, it allows the fitting of 3D profiles to the pixels that compose a reflection, the first two dimensions being the *xy* plane of the detector and the third the φ rotation. In combination, these should provide an optimum signal-to-noise ratio for the measured intensities and would appear to be the method of choice for data collection.

However, this involves a very large number of images, which can pose logistical problems in terms of data handling. Only if the read-out time is negligible in comparison with the exposure time can fine slicing be applied. If the detector read-out is slow, fine slicing becomes totally impractical. Multiwire chambers allow fine φ slicing, but unfortunately their disadvantages in terms of effective dynamic range preclude their use on high-intensity sources. Imaging plates are generally too slow for this approach.

The fine-slicing method is likely to receive a resurgence of interest with the introduction of fast read-out detectors such as the PILATUS, where diffraction images can be acquired continuously without the use of a beam shutter (Hülsen *et al.*, 2006).

9.1.6.7. Wide slicing

The object of the wide-slicing approach is to acquire the data on as small a number of individual exposures as possible. It involves large $\Delta\varphi$ values per image, usually of the order of 0.5° or more, which exceed the angular spread. Each image contains a considerable proportion of fully recorded reflections. Originally, wide slicing was used to minimize the large numbers of X-ray films to be processed. Only the wide-slicing approach is tractable for detector systems where the read-out time is relatively slow in relation to exposure, *e.g.* imaging plates with read-out times of 20 seconds to minutes.

Wide slicing has two drawbacks. Firstly, during integration of the intensity data, only 2D profiles are fitted for each individual spot in the wide slicing. Secondly, each reflection profile overlaps a background which accumulates throughout the whole time and

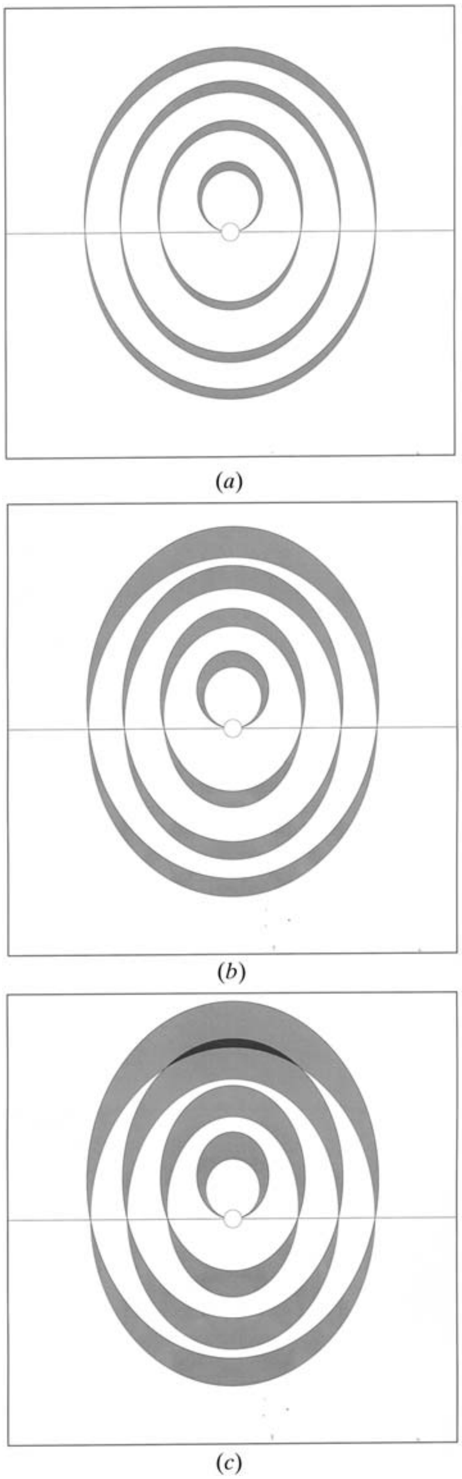


Figure 9.1.6.6
The width of the lunes is proportional to the rotation range per image, $\Delta\varphi$, which increases from (*a*) to (*c*). If the rotation range is large, the lunes overlap at high resolution.

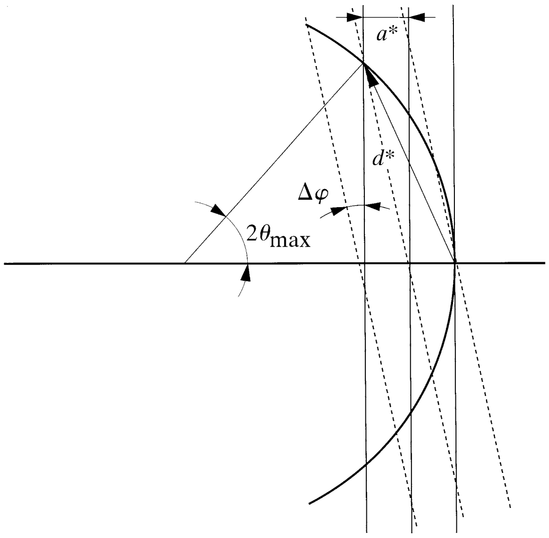


Figure 9.1.6.7
The largest allowed rotation range per exposure depends on the dimension of the primitive unit cell oriented along the X-ray beam; this is diminished by high mosaicity.

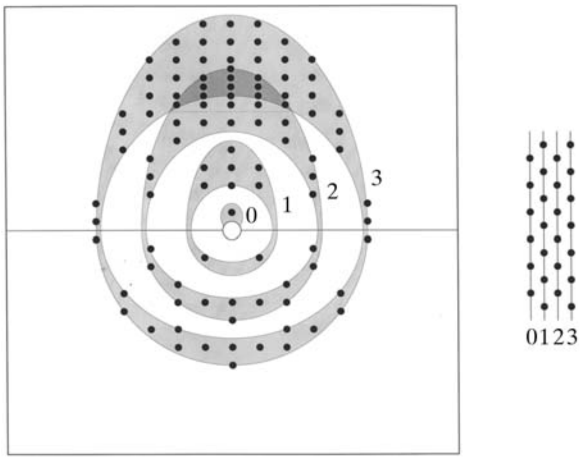


Figure 9.1.6.8
If the crystal lattice is centred or if its orientation is non-axial, the reflections do not overlap in spite of overlapping lunes, as illustrated on the right with consecutive layers of reflections viewed from the side.

angular range of the exposure, even when the reflection concerned is not diffracting.

The aim is to use the maximum acceptable rotation range per image. The lunes on an image have a finite width proportional to the rotation range. This width restricts the allowed angular range per image, as overlap of spots resulting from overlap of adjacent lunes must be avoided if the intensities are to be successfully integrated (Fig. 9.1.6.6). Several factors affect the degree of overlap and will be discussed in the rest of this section. A simple formula (Fig. 9.1.6.7) can be used to estimate the maximum permitted rotation range per image:

$$\Delta\varphi = 180d/\pi a - \xi,$$

where the factor $180/\pi$ converts radians to degrees, ξ is the angular spread of the reflection, d is the high-resolution limit and a is the length of the primitive cell dimension along the direction of the X-ray beam. However, this simplistic equation can be somewhat misleading. It most strictly applies when the lunes are densely packed with reflections, for an orthogonal cell rotated about a major axis. If this is not the case, then often rows of reflections from one lune fit between rows in the adjacent lune without overlap. For example, for a trigonal crystal with its *a* axis along the beam and rotating about its *c* axis, even and odd lunes contain rows of reflections that lie between one another on the detector (Fig. 9.1.6.8).

It can be extremely hard to record data from samples with a very long cell dimension. If the long axis lies along the X-ray beam, then it will restrict $\Delta\varphi$ considerably to very low values. This is exacerbated if the mosaicity is substantial. It is therefore beneficial to have the longest axis oriented roughly along the spindle axis, as it can then never lie parallel to the beam. This can be a problem with cryogenic samples mounted in loops, where the preferred orientation is hard to dictate, and this is an example where a κ-goniostat is an advantage, allowing reorientation of the crystal.

Simplistic application of the above equation for estimating $\Delta\varphi$ is not receommended. The degree of overlap also depends on pixel size, beam cross section, crystal size and mosaicity, and crystal-to-detector distance. Given the influence of these additional parameters, it is in practice better to employ the integration software first to interpret the diffraction pattern and to simulate predicted patterns by adjusting the data-collection parameters. Several modern packages have such strategy features, and it is vital to employ them before collecting data. In fact, from a theoretical point of view, the highest signal-to-noise ratio is obtained when $\Delta\varphi$ corresponds to the rocking curve of the diffraction, *i.e.* the mosaic spread plus the beam divergence.

9.1.7. Rotation method: geometrical completeness

This topic has been reviewed by Dauter (2005).

9.1.7.1. Total rotation range for non-anomalous data

The total set of structure-factor amplitudes from a crystal is a sphere of points in reciprocal space, with a radius defined by the maximum resolution. The intensities of the two hemispheres of data show a centrosymmetric relationship based on Friedel's law, which only breaks down if anomalous scatterers are present. However, the diffraction pattern possesses internal symmetry related to that of the real-space unit cell. This means that for all space groups an asymmetric unit of reciprocal space can be defined. Provided the intensities of all reflections in this asymmetric unit have been measured, those of all others can be generated by the symmetry operations and the Fourier transform for the complete structure computed.

The asymmetric unit has the shape of a wedge extending from the origin at the centre of the reciprocal sphere with a cutoff at a maximum radius corresponding to the limiting diffraction angle (resolution). Once the Laue symmetry group of the crystal has been determined (*International Tables for Crystallography*, Volume A, 2005), it is straightforward to define the shape of this wedge and establish which data must be recorded to make up a complete unique set. For macromolecular crystals, where there can be no centre of symmetry, the possibilities are further simplified to the point group rather than the Laue group. All space groups belonging to the same point group have the same asymmetric unit. The only differences lie in the presence or absence of screw axes or centring. Thus, space groups $P2_12_12_1$, $P2_12_12$, $P222_1$, $P222$, $I222$ and $I2_12_12_1$ all belong to point group (symmetry class) 222 and have the same asymmetric unit in reciprocal space. The only consequence of the presence of screw axes or lattice centring is to introduce systematic absences for some classes of reflection within this asymmetric unit of the point group.

It is usual to define the limits of the asymmetric unit by placing restrictions on the indices. For point group 222, the common conventional choice of limits on the reflection indices *hkl* is

$$0 \leq h \leq h_{\max}, \qquad 0 \leq k \leq k_{\max} \qquad 0 \leq l \leq l_{\max},$$

where $h_{\max}$, $k_{\max}$ and $l_{\max}$ are defined by the maximum resolution. In all point groups, there are multiple but equivalent ways of defining the asymmetric unit, but a default definition is generally chosen by the data-reduction software. For example, in triclinic symmetry, any hemisphere constitutes an asymmetric unit, and there are three typical choices of index limits:

$$0 \leq h \leq h_{\max}, \qquad \bar{k}_{\min} \leq k \leq k_{\max}, \qquad \bar{l}_{\min} \leq l \leq l_{\max},$$

or

$$\bar{h}_{\min} \leq h \leq h_{\max}, \qquad 0 \leq k \leq k_{\max}, \qquad \bar{l}_{\min} \leq l \leq l_{\max},$$

or

$$\bar{h}_{\min} \leq h \leq h_{\max}, \qquad \bar{k}_{\min} \leq k \leq k_{\max}, \qquad 0 \leq l \leq l_{\max}.$$

The standard choices of asymmetric unit taken from the *CCP*4 program suite (Collaborative Computational Project Number 4, 1994) are shown in Table 9.1.7.1.

The data are complete if the Ewald sphere has been crossed by all reflections in the asymmetric part of the reciprocal lattice. During data acquisition and reduction, all measured indices are conventionally transformed to this asymmetric unit of reciprocal space. Firstly, this allows merging of symmetry-equivalent measurements as appropriate. Secondly, it allows the completeness of the data to be assessed efficiently, using contributions from the whole sphere.

For all point groups, rotation of the crystal by 180° from any starting angle on the φ spindle axis is sufficient to provide a complete set of data (this is not sufficient if anomalous-dispersion measurements are required; see Section 9.1.7.2). Given such a total rotation, the redundancy of the measurements will increase with higher crystal symmetry. Thus, for a triclinic space group, the unique data will be measured almost twice on average (see the blind region below); for orthorhombic, eight times; for hexagonal class 6, 12 times; and for 622, 24 times. Redundancy is, in principle, advantageous, giving improved data quality (again see below), but it is generally possible to record complete unique data with a minimal overall rotation and correctly chosen starting angle on the spindle. It is of course necessary to determine the crystal orientation matrix, and this remains a vital part of the data-collection strategy. To minimize the effects of radiation

Table 9.1.7.1

Standard choice of asymmetric unit in reciprocal space for different point groups from the *CCP4* program suite

Point group	Index limits	
1	*hkl*: *hk*0: 0*k*0:	$l \geq 0$ $h \geq 0$ $k \geq 0$
2	*hkl*: *hk*0:	$k \geq 0, l \geq 0$ $h \geq 0$
222	*hkl*:	$h \geq 0, k \geq 0, l \geq 0$
4	*hkl*: 0*kl*:	$h \geq 0, k > 0, l \geq 0$ $k \geq 0$
422	*hkl*:	$h \geq k, k \geq 0, l \geq 0$
3	*hkl*: 00*l*:	$h \geq 0, k > 0$ $l > 0$
321	*hkl*: *hhl*:	$h \geq k, k \geq 0$ $l \geq 0$
312	*hkl*: *h*0*l*:	$h \geq k, k \geq 0$ $l \geq 0$
6	*hkl*: 0*kl*:	$h \geq 0, k > 0, l \geq 0$ $k \geq 0$
622	*hkl*:	$h \geq k, k \geq 0, l \geq 0$
23	*hkl*: *hkh*:	$h \geq 0, k > h, l > h$ $k \geq h$
432	*hkl*:	$h \geq 0, k \geq l, l \geq h$

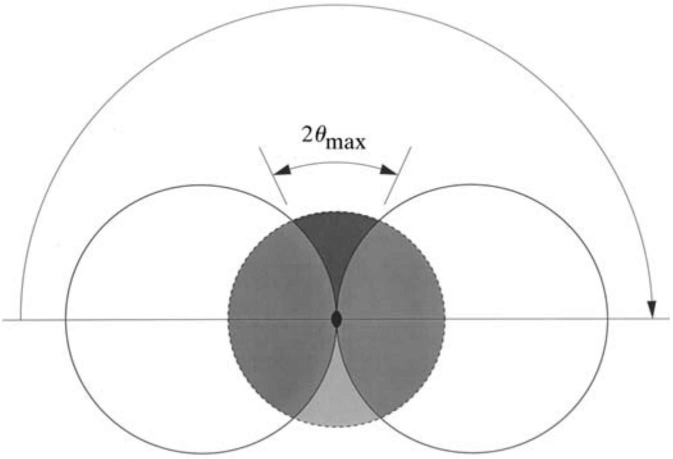


Figure 9.1.7.1
Rotation of a triclinic crystal by 180° in the X-ray beam, represented as rotating the Ewald sphere with a stationary crystal, projected along the rotation axis. For the purpose of analysing the relation of data completeness to crystal symmetry and orientation the two representations are equivalent.

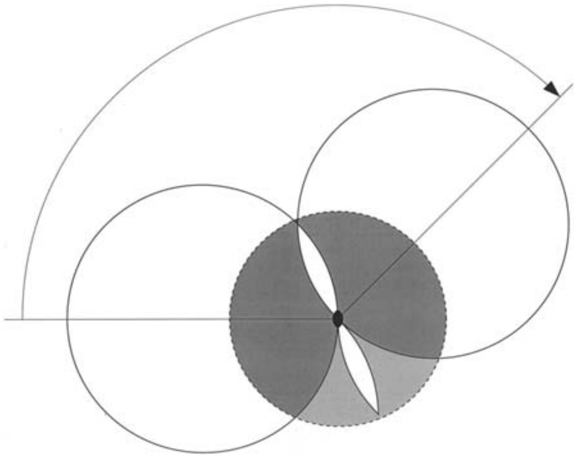


Figure 9.1.7.2
Rotation of a triclinic crystal by 135° is not sufficient to obtain totally complete data. At high resolution the completeness is higher than at low resolution, where a full 180° rotation is required.

damage, it is often beneficial to collect complete data with the minimal rotation range. This may well change with the advent of extremely fast detectors, when the decision-making process may take longer than data collection.

Thus, the crystal point-group symmetry has a profound effect on the total rotation range and the optimal starting spindle and crystal orientation for the most efficient recording of complete unique data. The rest of this section suggests strategies for the collection of complete data with minimal total rotation when anomalous-dispersion measurements are not required.

As stated above, for all crystals, rotation by 180° is sufficient to cover fully both sides of the Ewald sphere with intensity measurements. This is necessary for a triclinic crystal rotated around any arbitrary axis and also for a monoclinic crystal rotated around its unique *b* axis (Fig. 9.1.7.1). A twofold redundancy of unique data results; fourfold for the monoclinic case. Now consider a rotation of less than 180° (Fig. 9.1.7.2). Owing to the curvature of the Ewald sphere and the centre of symmetry arising from Friedel's law, the region of the sphere with reflections measured twice is diminished, and for part of the sphere there are no measurements. Most importantly, the proportions are resolution dependent. With a limited rotation, the high-resolution intensities reach a higher completeness than those at low resolution: data 90% complete at high resolution may be missing 20% of the low-resolution shells. Indeed, the low-resolution terms only become complete when a full 180° has been achieved.

The major data-processing software packages provide estimates of overall completeness as a function of total rotation range and starting point. However, they tend to neglect this variation with resolution. The fundamental importance of completeness at low resolution will be returned to later.

For total rotation by a given percentage of the angle needed to provide complete data, the resulting percentage completeness will be higher, again as a consequence of the curvature of the Ewald sphere. Consider again the triclinic case, when complete data require rotation by 180°. A single continuous range of 90° gives a completeness of about 65% (Fig. 9.1.7.3). Splitting the rotation range is advantageous; for example, if the crystal is rotated over two ranges of 45°, separated by a gap of 45°, the completeness typically rises to about 80%. In summary, for the triclinic case, the starting point and crystal orientation are irrelevant, but if it is impossible to cover 180° in the time available, it is better to use two or more sets of ranges. The software can again often provide advice on such strategies.

When the crystal has symmetry elements, the situation is more complex. Now the completeness is sensitive to the starting point of rotation and the crystal orientation, as well as the total rotation range used. All three must be considered in defining an optimum strategy for minimal rotation to give complete data. Consider an orthorhombic unit cell where the asymmetric unit comprises any octant of the reciprocal lattice. Minimal complete data require a total rotation of 90° between any twofold axis and the plane perpendicular to it (Fig. 9.1.7.4). This requires that one of the major axes must lie along the direction of the beam, either at the start or end of the 90° rotation, when the other two axes will lie in the plane of the detector. It is not necessary to rotate around one of the major axes, but the rotation axis should lie in one of the three major planes. If these conditions on crystal orientation or starting point are not satisfied, then more than a 90° rotation will be required. The proper selection of starting point is vital. A 90°

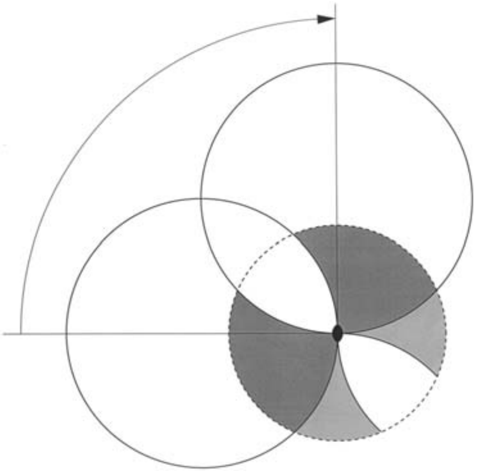


Figure 9.1.7.3
After a 90° rotation out of a required 180°, the overall completeness is higher than 50%.

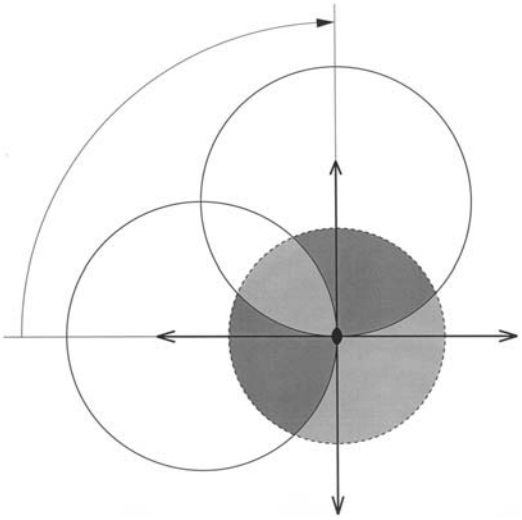


Figure 9.1.7.4
For an orthorhombic crystal, a 90° rotation is sufficient provided the starting or final orientation is along the major axis.

rotation starting midway between two axial positions, when the major axis only lies along the beam after 45°, will reduce the completeness after 90° to about 65%, since in essence the same 45° of unique data will be measured twice, albeit with high redundancy (Fig. 9.1.7.5). This emphasizes the need to define the crystal symmetry and orientation properly before data collection if minimalist protocols are to be employed.

In general, the higher the crystal symmetry, the more the completeness depends on the crystal orientation. In point groups 321 or 312, the asymmetric unit may be defined as a 30°-wide wedge that spans the space between the positive and negative direction of the threefold *c* axis. The index limits are

$$0 \le h \le h_{\max}, \qquad 0 \le k \le h, \qquad \bar{l}_{\max} \le l \le l_{\max}.$$

If the crystal is mounted with the threefold axis along the rotation spindle, it is sufficient to rotate by 30°, but only if the *a* or *b* axis lies along the beam at the start or end of the range. In contrast, if the crystal is rotated around **a** or **b**, then it is necessary to cover 90°. The second procedure will lead to a threefold increase in redundancy, but at the expense of a longer time.

The total rotation requirements for various crystal symmetries and orientations are given in Table 9.1.7.2. It is difficult to give reliable estimations for cubic crystals, since then the requirements vary dramatically with the crystal orientation.

Table 9.1.7.2
Rotation range (°) required in different crystal classes

The direction of the spindle axis is given in parentheses; *ac* means any vector in the *ac* plane.

Point group	Native data	Anomalous data
1	180 (any)	$180 + 2\theta_{\max}$ (any)
2	180 (*b*); 90 (*ac*)	180 (*b*); $180 + 2\theta_{\max}$ (*ac*)
222	90 (*ab* or *ac* or *bc*)	90 (*ab* or *ac* or *bc*)
4	90 (*c* or *ab*)	90 (*c*); $90 + \theta_{\max}$ (*ab*)
422	45 (*c*); 90 (*ab*)	45 (*c*); 90 (*ab*)
3	60 (*c*); 90 (*ab*)	$60 + 2\theta_{\max}$ (*c*); $90 + \theta_{\max}$ (*ab*)
32	30 (*c*); 90 (*ab*)	$30 + \theta_{\max}$ (*c*); 90 (*ab*)
6	60 (*c*); 90 (*ab*)	60 (*c*); $90 + \theta_{\max}$ (*ab*)
622	30 (*c*); 90 (*ab*)	30 (*c*); 90 (*ab*)
23	~60	~70
432	~35	~45

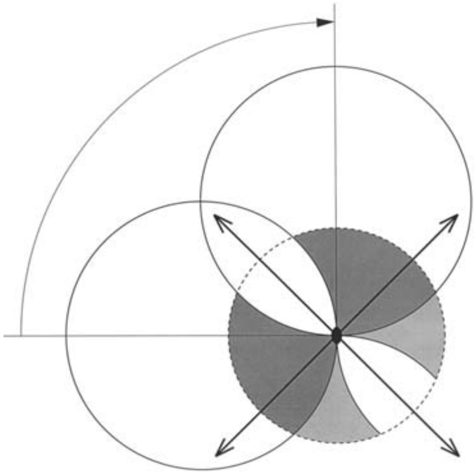


Figure 9.1.7.5
Rotation of an orthorhombic crystal by 90° between two diagonal orientations leaves a part of the reciprocal space unmeasured.

In the above, it was assumed that the detector was mounted centrally with respect to the incident X-ray beam. If it is offset either by a 2θ arm or by a translation, then the completeness for any total rotation range will be reduced. Software will generally be required to estimate the effective completeness and derive optimum strategies. For minimalist approaches to obtaining a high completeness, the importance of selecting the total rotation range, the optimal starting point and indeed the crystal orientation must be stressed. This means that the crystal orientation must be defined at the start of the experiment from the initial exposures.

9.1.7.2. *Total rotation range for anomalous-dispersion data*

In the presence of anomalous-scattering centres, Friedel's law breaks down and the intensities of the two halves of the reciprocal sphere are no longer equivalent. Strictly speaking, reflections related by a centre of symmetry or mirror relation cease to have equal intensities, but those related by pure rotation preserve their equivalence. The non-equivalent pairs of reflections are known as Bijvoet pairs. In macromolecular crystallography, it is often highly desirable to record the intensity differences between the Bijvoet mates to provide information on the position of anomalous scatterers, usually to be exploited in phasing procedures (Part 14). The anomalous signal should also be retained for so-called native data, for example in the discrimination between water and ions in the surface solvent shell.

This implies that the intensities of the unique reflections have to be measured for both hemispheres of reciprocal space. In the general (triclinic) case, this requires the rotation of the crystal by a wider rotation range. At very low resolution, the surface of the Ewald sphere can be approximated by a plane. In this case, rotation of the lower half of the Ewald sphere will cover a full hemisphere of data, and the upper half will cover the remaining centrosymmetrically related hemisphere. At high resolution, the surface of the Ewald sphere increasingly deviates from planarity by the Bragg angle θ on each side (Fig. 9.1.7.6). To record complete anomalous data for such a triclinic crystal therefore requires it to be rotated by $180° + 2\theta_{\max}$ from a random starting position. This will measure each Bijvoet mate at least once. However, only after a total rotation of 360° will the average multiplicity reach a value of two.

Similar reasoning applies to higher-symmetry space groups. Intensity data for two asymmetric units related by a centre of

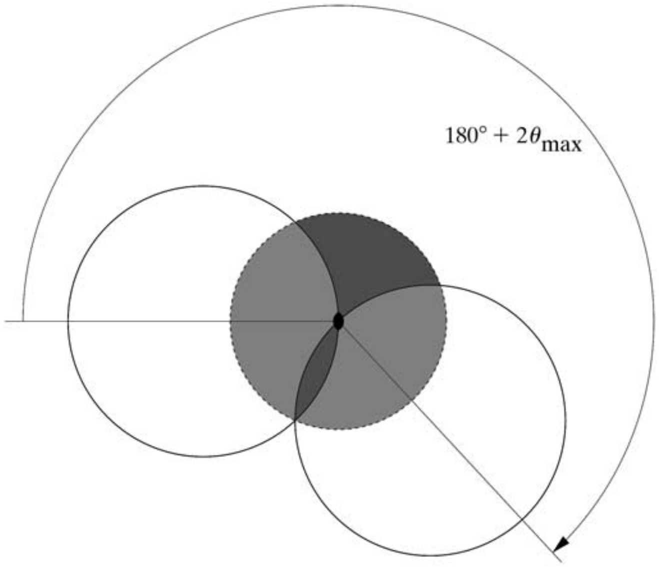


Figure 9.1.7.6
For data containing an anomalous signal, when both Bijvoet mates have to be measured, 180° rotation of a triclinic crystal is not sufficient and at least an additional $2\theta_{\max}$ is required.

symmetry or a mirror need to be recorded. For some cases, the total range remains the same for completeness of anomalous data as for native. However, in several symmetries or orientations, the total range must again be increased by either $\theta_{\max}$ or $2\theta_{\max}$ (Table 9.1.7.2).

9.1.7.3. Blind region

Even after rotation of the crystal about a single axis by 360°, some reflections do not cross the surface of the Ewald sphere and cannot be measured. These lie in a cusp around the rotation axis which is referred to as the blind region. This is in principle a disadvantage of the single-rotation method, but for most systems the problems are easily overcome. Owing to the curvature of the Ewald sphere, the width of the blind region increases with the resolution and directly depends on a single parameter, the diffraction angle θ (Fig. 9.1.7.7). The fraction, B_θ, of unrecordable reflections lying in the blind region at a particular resolution with Bragg angle θ is given by

$$B_\theta = 1 - \cos\theta.$$

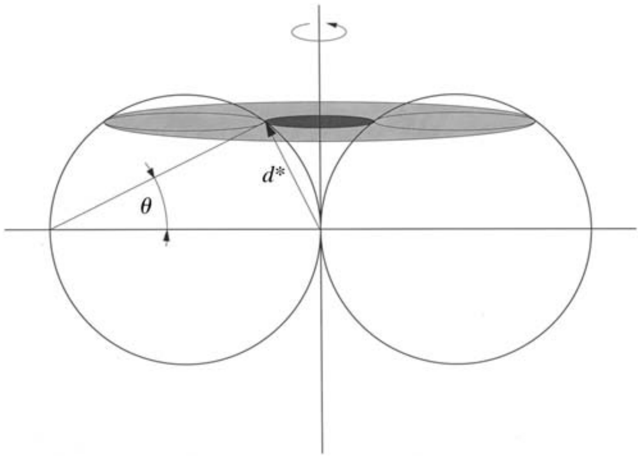


Figure 9.1.7.7
Rotation by 360° leaves the part of the reciprocal space in the blind region unmeasured, since the reflections near the rotation axis do not cross the surface of the Ewald sphere. The rotation axis in this projection lies vertically in the plane of the figure.

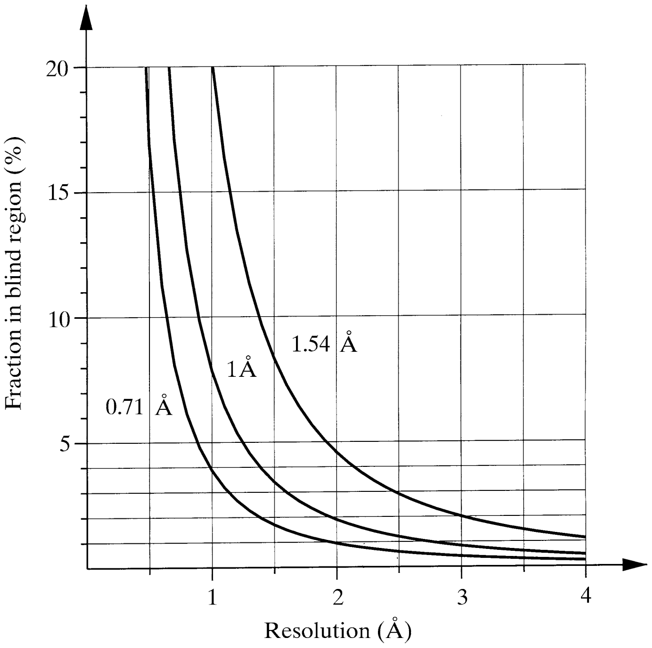


Figure 9.1.7.8
Dependence of the total fraction of reflections in the blind region on the resolution for three different wavelengths: 1.54, 1 and 0.71 Å.

The cumulative fraction, $B_{\rm tot}$, of reflections in the blind region up to a certain resolution is given by

$$B_{\rm tot} = 1 - 3(4\theta - \sin 4\theta)/(32\sin^3\theta).$$

$B_{\rm tot}$ is shown graphically as a function of resolution for selected wavelengths in Fig. 9.1.7.8.

For a particular resolution limit, the blind region is narrower if the wavelength is short, since the surface of the Ewald sphere is flatter (Fig. 9.1.7.9). This is an advantage of using short-wavelength radiation. For Cu $K\alpha$ radiation at 2.0 Å resolution, the blind region amounts to less than 5%. With shorter wavelengths, it falls below 2%.

The two halves of the blind region on either side of the Ewald sphere are related by the centre of symmetry. In the triclinic case, the blind region is therefore unavoidable with a single mount of

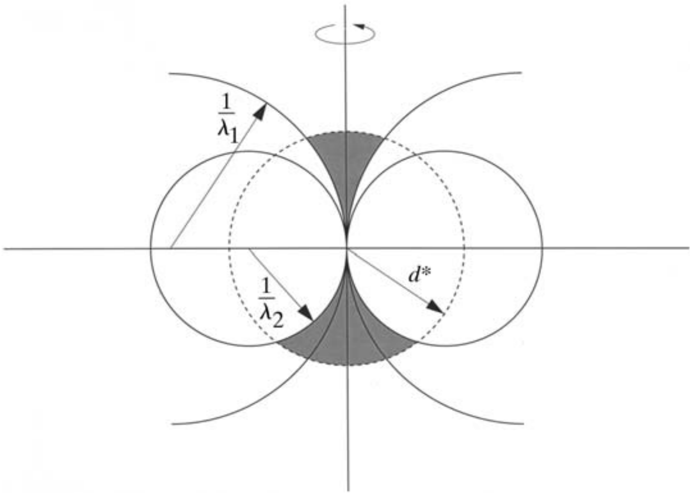


Figure 9.1.7.9
For shorter wavelengths the blind region is narrower, since the Ewald sphere is flatter.

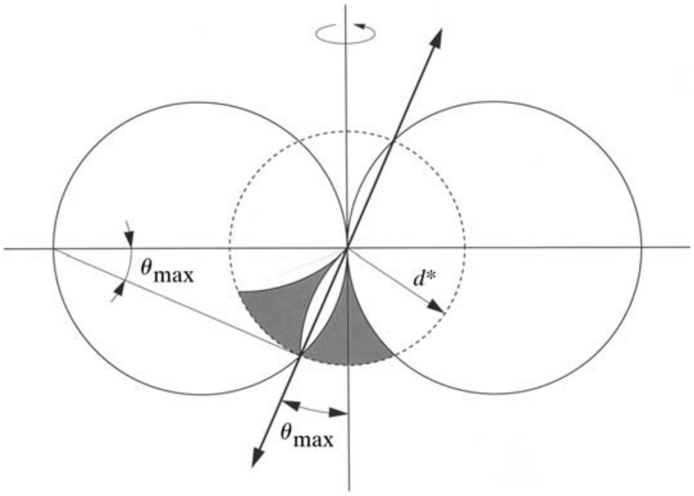


Figure 9.1.7.10
If the crystal has a symmetry axis, it should be skewed from the rotation axis by at least $\theta_{\max}$ to be able to collect the reflections equivalent to those in the blind region.

the crystal. The only solutions are to use a second mount of the crystal offset by at least 2θ from the first, easily achievable with a κ-goniostat, or to measure from a second sample.

For crystals with symmetry higher than $P1$, reflections that are symmetry-equivalent to those in the blind region may be recorded, and there will be no loss of unique reflections. Only if the unique axis passes through the blind region approximately parallel to the spindle axis will the reflections lying close to it not be repeated by symmetry in another region of reciprocal space. To avoid the blind region, it is sufficient to misorient the unique symmetry axis by at least $\theta_{\max}$ from the rotation axis (Fig. 9.1.7.10). To achieve full completeness, monoclinic crystals should not be oriented along the unique twofold axis or along any vector in the *ac* plane.

The reciprocal-lattice points on the border of the blind region cross the surface of the Ewald sphere at a very acute angle or fail to cross it completely, staying in the diffracting position for a considerable time. Their intensity cannot be measured accurately, because the Lorentz factor is large and its magnitude is very sensitive to minor errors in the orientation matrix. These reflections are located on the detector window along the line parallel to the spindle axis and should not be integrated.

The detrimental effect of the blind region on the completeness of data is negligible at medium and low resolution or if the crystal is non-axially oriented. This means that a simple single rotation axis is sufficient for the majority of applications.

9.1.7.4. *Alternative indexing*

If the crystal point-group symmetry is lower than the symmetry of its Bravais lattice, then the reflections can be indexed in more than one way. In other words, the symmetry of the reflection positions is higher than the symmetry of the distribution of their intensities. This situation typically arises for point groups with polar axes, such as groups 3, 4 or 6, which can be indexed with the *c* axis pointing in either one of two directions. The lattice does not define the directionality of such axes if its two remaining cell dimensions are equivalent. This problem does not occur in the monoclinic system, despite the polar twofold axis, as the two other axes are not equivalent. The most complex case is point group 3, which can be indexed in the 622 lattice in four non-equivalent ways. The other such groups have only two alternatives.

Table 9.1.7.3
Space groups with alternative, non-equivalent indexing schemes

Symmetry operations required for re-indexing are given as relations of indices and in the matrix form. In brackets are the chiral pairs of space groups indistinguishable by diffraction. These space groups may also display the effect of merohedral twinning, with the twinning symmetry operators the same as those required for re-indexing.

Space group	Re-indexing transformation	
$P4$, $(P4_1, P4_3)$, $P4_2$, $I4$, $I4_1$	$hkl \rightarrow kh\bar{l}$	$010/100/00\bar{1}$
$P3$, $(P3_1, P3_2)$	$hkl \rightarrow \bar{h}\bar{k}l$	$\bar{1}00/0\bar{1}0/001$
	or $hkl \rightarrow kh\bar{l}$	$010/100/00\bar{1}$
	or $hkl \rightarrow \bar{k}\bar{h}\bar{l}$	$0\bar{1}0/\bar{1}00/00\bar{1}$
$R3$	$hkl \rightarrow kh\bar{l}$	$010/100/00\bar{1}$
$P321$, $(P3_121, P3_221)$	$hkl \rightarrow \bar{h}\bar{k}l$	$\bar{1}00/0\bar{1}0/001$
$P312$, $(P3_112, P3_212)$	$hkl \rightarrow \bar{h}\bar{k}l$	$\bar{1}00/0\bar{1}0/001$
$P6$, $(P6_1, P6_5)$, $(P6_2, P6_4)$, $P6_3$	$hkl \rightarrow kh\bar{l}$	$010/100/00\bar{1}$
$P23$, $P2_13$, $(I23, I2_13)$, $F23$	$hkl \rightarrow k\bar{h}l$	$010/\bar{1}00/001$

There is an analogous problem for cubic space groups within point group 23. Here the lattice possesses fourfold symmetry, but the intensity distribution has only twofold symmetry. Rotation by 90° leads to alternative, although perfectly permitted, indexing of reflections.

Each allowed scheme is permitted and self-consistent for a single crystal, since all possibilities will perfectly match the crystal lattice. However, under alternative indexing schemes, the same reflection will be given different indices, which can pose problems when data from more than one crystal are to be merged or compared. Merging is needed when more than one sample is required to record a complete data set. Comparison is needed when looking for heavy-atom derivatives or for ligand complexes with isomorphous crystals. For these, the reflections of one crystal must be selected as a standard, and it is easy to make other crystals consistent with this standard either by changing the orientation matrix at the time of intensity integration or by applying re-indexing to the integrated intensity set. The alternative indexing schemes are related by those symmetry operations present within the higher symmetry of the Bravais lattice but absent from the point-group symmetry. The point groups with alternative indexing systems are shown in Table 9.1.7.3, together with the necessary symmetry operations for re-indexing.

Several experiments require the recording of multiple data sets from the same crystal. One example is the collection of more than one pass with different exposure times (see below), and a second is in multiwavelength anomalous dispersion (MAD) experiments. In these experiments, the software systems may independently choose any of the alternative systems for different sets, which may then be incompatible and need re-indexing. It is much simpler to ensure a common orientation matrix modified as appropriate for all sets at the time of intensity integration.

9.1.8. Crystal-to-detector distance

The crystal-to-detector distance (CTDD) should be selected so that the whole area of the detector is usefully exploited. The shorter the CTDD, the higher the resolution of the indexed reflections at the edge of the image; but if the CTDD is too short, then the outer regions of the detector window record only indices with attached noise rather than intensities. A longer CTDD spreads the background radiation over a larger area of the detector as the background level diminishes in proportion to the square of the CTDD. In contrast, owing to collimation and focusing, the profiles of the Bragg reflections do not broaden so much, and the signal-to-noise ratio is enhanced at longer distances. It is advantageous to use the largest possible CTDD under the condition that meaningful data extend to, but not beyond, the active edge of the detector.

It is not straightforward to judge the resolution limit of meaningful diffraction. The most scientific approach involves recording, processing and merging a small number of images and making a decision on the basis of the resulting intensity statistics. However, this does require time, which should only pose a problem on ultra-high-intensity sources with very rapid data collection. A more pragmatic approach relies on visual inspection of the initial exposures using a graphical display at various contrast levels. Normally, if reflections are not visible by eye at the highest display contrast, their intensities are not meaningful. Some safety margin can be applied by setting the CTDD to a slightly shorter value than that estimated from visual inspection. Naturally, the resolution limit to which meaningful intensities extend depends on the exposure time, and the decision concerning the CTDD should follow the selection of the appropriate exposure (Section 9.1.11.2).

In addition to the significance of the reflection intensities, another important factor is the spatial resolution of spot profiles on the detector. If the crystal cell dimensions are large, the profiles may superimpose and the reflections may be impossible to integrate. At longer CTDD, the diffraction pattern spreads out and the profile overlap diminishes. If necessary, the detector can be offset from the central position to measure high-resolution data at long CTDD, but a larger total rotation is required to reach full data completeness. This applies only if the overlap of profiles belonging to the same lune results from a long axis lying parallel to the detector plane. The superposition of reflection profiles resulting from overlapping lunes will not be alleviated by increasing the CTDD; the only remedy for this is to reduce the rotation range $\Delta\varphi$ per exposure.

In addition to the proper selection of the CTDD, attention should be paid to the proper positioning of the beam stop. It should be centred with respect to the direct beam and cover the beam cross section completely. No part of the direct beam should reach the detector, and there should be no indirect scatter by the beam stop. The optimal reduction of air scatter is to have the smallest beam stop consistent with the dimensions of the beam, placed as close to the crystal as possible. For a given size of beam stop, the crystal-to-beam stop distance should be matched to the CTDD, sufficiently far from the crystal to minimize its shadow and concomitant obstruction of the valuable lowest-resolution reflections. If the beam stop is mounted on a metal wire, it is better to position the wire along the spindle axis where it will only interfere with those reflections around the blind region.

9.1.9. Wavelength

The wavelength of X-radiation can be tuned only at synchrotron sources. Rotating-anode generators produce radiation at a fixed wavelength which is characteristic of the metal of the anode, usually copper with $\lambda = 1.542$ Å.

The proper selection of the wavelength is most important for collecting data containing an anomalous-scattering signal. In general, the imaginary component $\Delta f''$ of the anomalous-dispersion signal is high on the short-wavelength side of the absorption edge of the anomalous scatterer present in the crystal. Near the absorption edge, both components, real $\Delta f'$ and imaginary $\Delta f''$, vary significantly. This variation is utilized in the MAD technique, the strict requirements of which are discussed in Chapters 14.2 and 14.3.

If the data are collected using a single wavelength with the aim of measuring Bijvoet differences, $\Delta F_{\rm anom} = F^{+} - F^{-}$, the requirements are not as strict as for MAD. However, it may be advisable to record the fluorescence spectrum around the region of the expected absorption edge. If the fluorescence signal from the crystalline sample is too weak, the appropriate metal or salt standard can be used. When using anomalous scatterers displaying large white lines within their spectra, the wavelength should be accurately adjusted on the basis of the spectrum measured from the actual sample.

For collecting data without an anomalous signal, there are no strict requirements concerning the wavelength. The maximum intensity provided by the beamline depends on the energy of particles in the synchrotron storage ring and on the beamline optics. Typically, wavelengths around 1 Å or shorter are used at most synchrotrons, assuring high beam intensity and low absorption of X-rays by the sample and air, thus reducing the radiation damage of the crystal. This is of particular importance at the very bright beamlines at third-generation synchrotrons. To diminish the effect of air absorption further, it is possible to fill the space between the crystal and the detector with helium. Short wavelengths are advantageous for collecting high-resolution data, since the diffraction angles are smaller and there is no need to use a very short CTDD. The effect of profile elongation owing to the oblique incidence of diffracted X-ray beams on the detector is then smaller, and the blind region is narrower.

9.1.10. Lysozyme as an example

Tetragonal hen egg-white lysozyme (Chapter 25.1 and Blake *et al.*, 1967), crystallizing in the space group $P4_32_12$ with cell dimensions $a = b = 78.6$ and $c = 37.2$ Å, is used here as a model system to illustrate some of the points made above, based on Dauter (1999). The example involves a set of two consecutive blocks of images with a crystal-to-detector distance of 243 mm, a wavelength of 0.92 Å, a resolution of 2.7 Å, an oscillation range of 1.5° and a crystal mosaicity around 0.5°. These images are shown in Fig. 9.1.10.1(*a–f*).

The first four images, (*a–d*), were exposed with the tetragonal fourfold *c* axis lying approximately along the direction of the beam. On these images, the reflections within each lune are arranged in a square grid, reflecting the tetragonal symmetry with $a = b$. The squares are oriented with their diagonals in the horizontal and vertical directions of the image, as the crystal was mounted with its [110] direction along the spindle rotation axis. Indeed, at the end of image (*a*) and the start of image (*b*), the *c* axis lay almost perfectly along the beam, and the zero-layer lune almost disappears behind the beam-stop shadow, since the corresponding (*hk*0) plane in reciprocal space is tangential to the Ewald sphere at the origin of the reciprocal lattice.

The lunes are widely spaced with clear gaps between them, because the third cell dimension, *c*, which is perpendicular to the detector plane, is relatively short, 37.2 Å. Images (*e–f*), exposed at an angle on the rotation spindle roughly 90° away from (*a–d*), have a quite different appearance, despite the rotation range per image being the same. Each lune is less densely populated by reflections, but the number of lunes is larger and the gaps between them much smaller. This arises from the lunes now being parallel to the (*hhl*) family of planes, as the $[1\bar{1}0]$ vector is now parallel to the beam. The interplanar spacing within this family is less than for those on images (*a–d*), hence at high resolution,

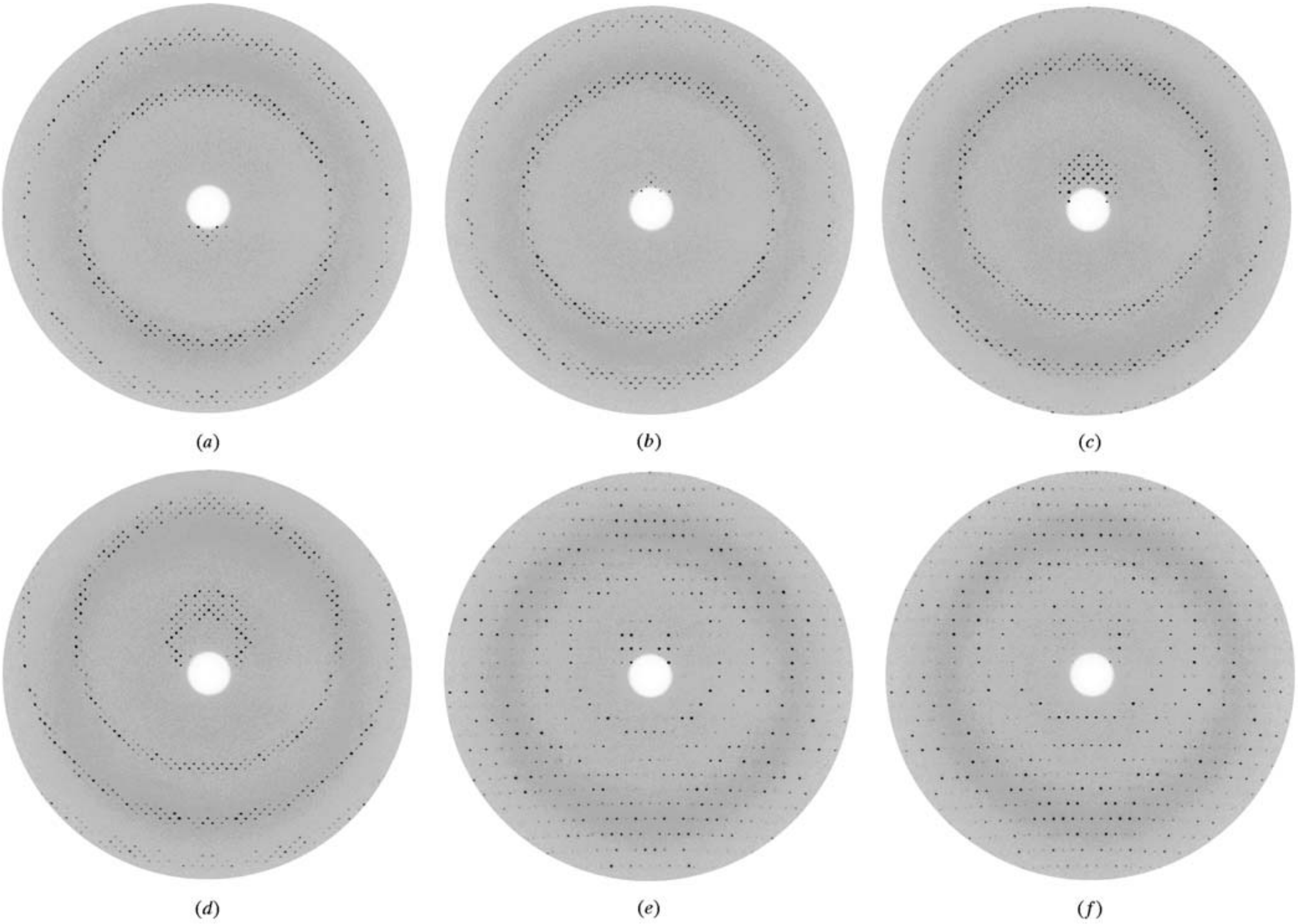


Figure 9.1.10.1
Images recorded from a crystal of lysozyme. (*a*–*d*) Four consecutive exposures with the crystal fourfold axis parallel to the X-ray beam. (*e*–*f*) Two successive exposures 90° away, when the fourfold axis lies vertically in the plane of the image. The crystal [110] direction is parallel to the rotation axis, horizontal in the plane of the images.

close to the edge of the detector window, the lunes overlap on images (*e*–*f*). The reflections, however, do not overlap, as the crystal orientation is diagonal; the lunes are sparsely populated, with large separation between adjacent spots, so the reflections on successive lunes fit between one another. It should be noted that the density of reflections in different regions of the reciprocal lattice is constant, and that the total number of reflections recorded on an image depends only on the rotation range, not on the crystal orientation.

The zero-layer lune containing reflections with indices $hk0$ is especially evident on exposures (*c*–*d*) directly above the centre of the image. With such a lune close to the centre, the reciprocal lattice shows minimal distortion owing to its projection onto the detector plane, and the lune appears as a 'pseudo-precession' pattern. The systematic absence of every second reflection, with odd index, along the $h00$ and $0k0$ lines indicates the presence of twofold screw axes of symmetry along the crystal axes a and b. Images (*e*–*f*), 90° away, have the hhl lune at the centre and, although it is less well separated from higher lunes, the presence of a fourfold screw axis along $\mathbf{c}$ is confirmed by the presence of only every fourth reflection on the $00l$ line. This allows the identification of the space group as $P4_12_12$ or its enantiomorph, $P4_32_12$. In general, the positions of the reflections define only the Bravais lattice, and it is the symmetry of the intensity pattern which reflects the point group. Thus, further confirmation that the symmetry belongs to point group $P422$ rather than $P4$ comes from the symmetric relation of the intensity distribution on either side of each lune in images (*a*–*d*). This is equivalent to the earlier use of precession photography for space-group elucidation.

Close inspection shows that the reflections at the edges of the lune are also present on the adjacent image. The rotation range was 1.5°, and the mosaicity was estimated at 0.5°, and thus about one-sixth of the reflections are partially recorded at each edge of the lune, giving one-third partially recorded terms in total. The lack of sharpness at the edge of the lunes confirms a substantial level of mosaicity.

9.1.11. Rotation method: qualitative factors

9.1.11.1. Inspection of reflection profiles

Reflection profiles should be checked on the first recorded images. Very often a quick inspection of the profiles can disqualify a bad crystal without further loss of time. The profiles should have a single maximum and smooth shoulders. If the crystal shape is irregular, it may be reflected in the spot profile. Profiles should not have double maxima or be substantially elongated or smeared out, which usually arises from crystal splitting. The profiles should certainly be inspected if initial autoindexing of the diffraction pattern is unsuccessful.

Even if the spot profiles appear to be regular on the first image, it is good practice to inspect a second image at a substantially different φ rotation angle, preferably 90° away, since crystal splitting may have a similar effect on the appearance of the lunes and profiles as does high mosaicity on a single image (Section

9.1.6.3). High mosaicity and splitting (often incorrectly referred to as twinning) must not be confused. If two parts of a split crystal are slightly rotated with respect to one another around a certain axis, the diffraction patterns will look different depending on the orientation. When such an axis is perpendicular to the detector plane, the spots will be doubled or smeared out. When the axis is parallel to the detector plane, the profiles resulting from the two parts of the crystal will overlap almost perfectly, but the lunes will be broadened, similar to the effect of high mosaicity.

After indexing the diffraction pattern, the integration profiles should be matched with the size and shape of the diffraction spots. The spots should not extend into the area defined as background. Selection of integration profiles that are too small will lead to incorrect integration of intensities. In contrast, if the profile areas are too large then the standard uncertainties will be wrongly estimated.

9.1.11.2. Exposure time

According to the principles of counting statistics, the longer the exposure, the better the signal in the data. The standard uncertainty of the measurement is equal to the square root of the number of counts, and the signal-to-noise ratio increases with the accumulated counts. In practice there are limitations to this rule.

The dynamic range and saturation limit of the detector is one limiting factor. It may be impossible to measure adequately the strongest as well as the weakest reflection simultaneously, since their intensities differ by several orders of magnitude. If the exposure time is long enough to record the weakest intensities, then in general at low resolution the most intense reflections may saturate some pixels within their profile on the detector. Such reflections are termed ‘overloads’ and this problem will be addressed in Section 9.1.11.3.

Exposure time can be limited by the total time available for the experiment. This is often a particularly acute problem for synchrotron-data collection, with high oversubscription of beamlines. The decisions concerning exposure time depend on the expected application of the data, since different applications have different requirements, as addressed in Section 9.1.13. Within the given time constraints, the first priority should be data completeness, even at the expense of underexposure. In this context, it is useful to recall that to increase the statistical signal-to-noise ratio by a factor of two, it is necessary to prolong the exposure time by at least a factor of four.

9.1.11.3. Overloads

Some detectors, or their associated read-out systems, are limited in the number of counts they can accumulate in one pixel. The number recorded reaches a maximum number which cannot be further increased, *i.e.* the pixels can become saturated. This means that these pixels retain the same maximum value on longer exposure whilst other, non-saturated, pixels continue to accumulate counts. The intensity in saturated pixels will hence be underestimated compared to the others and any intensities estimated from profiles including such pixels will be biased towards low values. It is essential that pixels that are saturated are flagged and recognized by the processing software. There are several ways to deal with the problem of saturation.

(1) Reject all reflections that contain saturated pixels. These will tend to be at low resolution. If more than a very few are rejected, this can be a truly disastrous choice, especially if the data are to be used for molecular replacement. In addition, missing the largest terms degrades the continuity and information content of all electron-density maps derived therefrom. This point is relevant to several applications (Section 9.1.13).

(2) Reject only those pixels that are saturated, and fit average standard profiles estimated from the non-saturated spots. This gives a poorer estimate than if the pixels were not saturated, but for applications such as molecular replacement or direct methods where the high-intensity data are essential it is certainly better than option (1).

(3) Reduce the exposure time to ensure that there are no overloaded pixels. This is a trade-off, because if there is a large contrast between the intensity of the weakest and the strongest terms in the pattern, then the weaker terms will have a low and possibly unacceptable signal-to-noise ratio under this regime.

(4) Use more than one pass through the rotation range, with different exposure times. The longest exposures should be sufficient to ensure that the intensities of the data at the high-resolution limit of the pattern are statistically significant. The shortest should ensure that the number of saturated pixels in the ‘low-resolution’ pass is minimized. If the contrast between the low- and high-resolution passes is too great, differing by a factor of much more than about ten, then additional passes with intermediate exposure times should be used to allow satisfactory scaling of the data from these images. The CTDD for each pass with shorter exposure should be increased only so as to cover the resolution to which reflections were saturated on the previous pass. The rotation range on individual images can then be increased accordingly, in the wide φ-slicing option. On bright synchrotron beamlines, if the second pass requires exceedingly fast rotation of the spindle-axis motor and rapid opening and closing of the beam shutter beyond the limit of reliability, it may be better to attenuate the beam, for example with a series of aluminium foils. As discussed in Section 9.1.7.1, if high-resolution data are collected in several passes with different exposures and resolution limits, it may not be necessary to cover all of the theoretically required rotation range in the highest-resolution pass. The curvature of the Ewald sphere results in the high-resolution data being completed with a smaller total rotation range than the low. It is vital that the lowest-resolution pass covers the total rotation range required for complete data.

Clearly, the optimum solution is to have a detector with a sufficient dynamic range to cover pixels of both weak and strong reflections. The dynamic range has already been increased with recent imaging plates and CCDs. Enhanced dynamic range may prove to be the most important advance of solid-state pixel detectors.

An additional advantage of the fine-slicing approach is that it leads to fewer overloads. Each reflection profile is divided between several separate images and as a result the effective dynamic range of the detector is increased.

9.1.11.4. R factor, $I/\sigma(I)$ ratio and estimated uncertainties

It is customary to judge data quality by the overall $R_{\rm merge}$, calculated using the squares of the structure-factor amplitudes (often erroneously called intensities):

$$R_{\rm merge}=\sum_{hkl}\sum_{i}|I_{hkl,i}-\langle I_{hkl}\rangle|/\sum_{hkl}\langle I_{hkl}\rangle.$$

$R_{\rm merge}$ provides a measure of the distribution of symmetry-equivalent observed intensities around the average value. However, the most popular form of $R_{\rm merge}$ given above is not a proper, statistically valid quantifier. It does not take into account the multiplicity of the measurements and, as a consequence, it actually rises with increased multiplicity, falsely indicating degradation of the data quality when in reality they have a higher accuracy. Modifications of $R_{\rm merge}$ have been proposed to include the effect of multiple measurements properly (Diederichs & Karplus, 1997; Weiss, 2001) (see Chapter 2.2).

A better quantity for assessing the quality of the X-ray data is the $\sum_{hkl} I_{hkl}/\sum_{hkl}\sigma(I_{hkl})$ ratio, provided the standard uncertainties, $\sigma(I)$, are correctly estimated. Detectors such as imaging plates or CCDs do not measure individual X-ray quanta directly, having a gain factor dependent on the response of the individual detector pixel to a single X-ray photon. If the gain factor is not known accurately for a particular detector, the resulting standard uncertainties of the measured intensities will be estimated at an incorrect level. If the multiplicity of the reflections is higher than unity, it is possible to correct the uncertainties *a posteriori*. This can be done either from a comparison with the expected values using the χ^2 test, or by using the *t*-plot. The latter requires that the ratio of the differences between equivalent intensity measurements to their standard uncertainties, $t = (I_i - \langle I \rangle)/\sigma(I_i)$, follows a normal distribution with a mean of 0.0 and standard deviation of 1.0. Both of these methods assume the errors have a normal distribution, and that only the mean and width have been incorrectly estimated and should be appropriately adjusted. They cannot take into account systematic errors of measurement.

The data-merging procedure in addition allows the identification of statistical 'outliers' and their exclusion from the data (Read, 1999). Outliers are defined as those observations that lie sufficiently far from the mean of a set, and assumption of a normal distribution suggests they suffer from substantial systematic errors of measurement. In a crystallographic experiment, outliers are those intensity measurements that deviate unexpectedly from the mean intensity of a set of symmetry-equivalent reflections. In the recording of rotation data, one typical source of such systematic errors is erroneous classification of reflections predicted as partially or fully recorded. This is a severe problem for those reflections lying close to the blind region. A second example is the presence of so-called 'zingers' in individual CCD detector pixels caused by scintillations from trace radioactivity of the taper glass. Other problems such as shadowed or inactive regions of the detector window give rise to a range of such systematic errors.

A small number of outliers may be expected from such causes. However, the total fraction of reflections flagged as outliers and rejected from the merging process should be small, certainly much less than 1%. Larger fractions indicate serious deficiencies in the hardware or the software and suggest something is very wrong with the experiment. There should always be a physical reason for rejecting outliers, other than just a need to reject those agreeing poorly with their symmetry-equivalent intensities in order to drive down $R_{\rm merge}$. It is always possible to reduce $R_{\rm merge}$ and to provide an apparent 'improvement' in the data by rejecting a large percentage of measurements, but this is extremely bad practice.

Good crystallographic data depend strongly on an appropriate statistical procedure. It is also inappropriate to exclude those reflections with intensities lower than a cutoff limit, such as 1σ, before or during the process of data merging. Weak intensities also carry information and their neglect introduces bias into the measured intensity distribution, affecting, for example, the overall or individual atomic temperature factors.

The true outer resolution limit of the diffraction pattern is not trivial to define and indeed depends to some extent on the application. If $I/\sigma(I)$ is higher than 1.0, then a resolution shell of data indeed contains some information in a statistical sense – provided of course that $\sigma(I)$ has been correctly estimated. However, as $I/\sigma(I)$ falls close to unity there will in practice be very few significant observations amongst a great deal of noise. It is necessary to make some decision about where to cut the effective resolution. For the application of direct methods, for example using *SHELXD* (Sheldrick, 2008), the cutoff is often defined as the resolution shell where $I/\sigma(I)$ falls to 2.0, when $R_{\rm merge}$ usually reaches 20–40% depending on the symmetry and redundancy. Cruickshank (1999*a*,*b*) has provided a formula for a data precision indicator (DPI) which includes the effect of falling $I/\sigma(I)$ ratio (see Chapter 2.2).

For other applications it may be advisable to accept even very weak data. Direct methods use only a subset of the most meaningful reflections, but which should extend to as high a resolution as possible. In addition, when the data are sparse from crystals that only diffract to very limited resolution, perhaps around 3 Å, then it is essential to retain all the experimental data, even if they are weak.

9.1.12. Radiation damage

9.1.12.1. Historical perspective

All crystals irradiated with X-rays absorb at least a fraction of the radiation, resulting in damage to the sample (Henderson, 1990). The energy from the absorbed photons may initially result in the disruption of chemical bonds, before being eventually dissipated as thermal energy. For well ordered small-molecule crystals the lattice is close-packed and the effects arising from the absorbed photons are restricted to the immediate environment of the absorption event, so-called primary damage. Only when a substantial fraction of the crystal has been affected do cooperative effects set in.

In contrast, roughly 50% of a macromolecular crystal is disordered aqueous solvent (Matthews, 1968). At room temperature this allows a secondary mechanism of radiation damage, resulting from diffusion of radicals and ions produced at the primary absorption site which affect chemical moieties at positions remote from this site. The details of this process remain poorly understood but are related to the extremely damaging effects of X-rays on biological tissue. A consequence of this damage is that degradation of the crystal order continues even after the irradiation is stopped or interrupted. For collection of data at room temperature from protein crystals mounted in capillaries, secondary damage contributes significantly to the rate of deterioration of the diffraction pattern. One of the gains of the early applications of SR was that it allowed recording of data to proceed ahead of the effects of secondary damage, increasing the effective, if not the absolute, lifetime of the crystal in the X-ray beam. An experiment often required several crystals, all of which showed the effects of temporal decay in their recorded intensities, which needed to be merged to provide complete data.

9.1.12.2. Cryogenic vitrification

In the early 1990s, the introduction of protein-data collection at cryogenic temperatures, using so-called flash cooling, was a major breakthrough (Garman & Schneider, 1997; Rodgers, 1997;

Garman & Owen, 2006). Such vitrification of crystals largely prevented the effects of secondary damage. On the X-ray sources then available, it was in most cases possible to record complete data from a single sample without significant degradation of the diffraction, enormously simplifying the strategy of data collection and merging.

Almost all data are currently collected from vitrified samples (see Part 10). The prolonged life of the sample and modest rates of data acquisition, even at second-generation SR sources with imaging plates, allow enough time for careful analysis of the initial images and optimization of the strategy.

A second major advantage of cryogenic data collection is that it allows crystals to be reused after initial data have been recorded. Two examples show the usefulness of this approach. Firstly, when screening the binding of heavy atoms for phase determination or ligands for complex formation, data can first be recorded to the minimum resolution needed to determine whether the binding is successful. Secondly, a series of vitrified crystals can be screened for their degree of order in the home laboratory, and the best stored and retained for subsequent improved collection either in the home laboratory or at a synchrotron site. The ability to transport vitrified crystals has proved invaluable in this respect, and leads to optimal use of synchrotron resources.

9.1.12.3. High-intensity third-generation SR sources

The advent of third-generation SR sources and insertion devices has led to X-ray beams of unprecedented intensity. The speed of data collection can be of the order of 1 second per 1° rotation. In association with CCD detectors able to read out images within a few seconds, this means that a complete data set can be obtained in a few minutes. At first sight, this would seem to have solved the problem of macromolecular data collection, as such speeds should allow recording of highly redundant accurate data to the highest resolution in a tractable time.

However, with such high intensities it appears that the effects of radiation damage are significant and result in specific effects on susceptible parts of the structure. The useful active exposure lifetime of typical crystals seems to be around five minutes, with substantial degradation of the diffraction pattern ensuing even for vitrified crystals. The first manifestation of radiation damage is the disruption of disulfide bridges and decarboxylation of aspartates and glutamates. This effect means appropriate strategies for selecting the optimal starting point of rotation in order to minimize the total rotation required for collection of complete data are once more essential. Several strategy programs, such as *BEST* (Popov & Bourenkov, 2003; Bourenkov & Popov, 2006), now permit this to be done effectively.

9.1.12.4. Correcting data for the effects of radiation damage

The overall effect of radiation damage is that the higher-resolution intensities decrease faster than those at low resolution. This effect is largely taken into account by the relative *B* factors applied to individual images during data scaling and merging by the major data-reduction programs.

However, such scaling does not allow for the effect of specific structural damage (*e.g.* the S–S bridges and carboxylic groups) on individual reflection intensities. A method to deal with this has been proposed by Diederichs *et al.* (2003). This is based on a zero-dose extrapolation of intensities and requires that a timestamp be attached to each individual intensity measurement. Such a timestamp has also been used to assist in estimation of phases by *SHARP* (Schiltz *et al.*, 2004).

9.1.13. Relating data collection to the problem in hand

The data-collection protocol should be matched to the purposes for which the data are to be used. Different applications present a range of different needs, requiring the intensities (or structure-factor amplitudes) to be exploited in different ways. In this section a representative set of applications is outlined in terms of how the tactics and strategies of data collection can vary.

9.1.13.1. Isomorphous-anomalous derivatives

The phasing of proteins by isomorphous replacement requires the collection of data from crystals of one or more heavy-atom derivatives of the protein that are isomorphous to the parent native crystal. Preparation of derivatives involves either soaking of native crystals in the heavy-atom solution or co-crystallization with the heavy-atom reagent (Part 12). Data collection can be split into two parts. The first step is to establish whether a potential derivative is isomorphous and contains the expected heavy atoms. The second is to collect the data on this derivative to provide the necessary phase information for the native structure factors. The problems of how to utilize the phase information are addressed in Part 12. Here, strategies applicable to the two steps are described.

Screening of derivatives can be carried out by collecting data to the resolution limits of the crystals. This can consume substantial data-collection resources and lead to irrelevant data that are not from isomorphous crystals or do not contain the anticipated heavy-atom signal. It is preferable to record the minimum data sufficient to identify a potential derivative in order to save time and resources, as many samples may need to be screened. A minimal strategy can exploit some or all of the following protocols:

(1) An essentially complete native-data reference set should be available, although not necessarily to the ultimate resolution limit.

(2) Preparation of a set of crystals with a selected set of potential heavy atoms, the number depending on crystal availability.

(3) Collection of a small number of images from each potential derivative crystal, ideally on the home-laboratory rotating-anode source or an SR beamline if necessary. These data can be recorded to a low resolution: in principle 4 Å or less should be enough. The resulting partial derivative data are scaled with the complete native set. The fractional isomorphous difference can be evaluated easily and compared with the expected agreement with the native data. In general, values less than 10% suggest that the heavy atom is not bound. Values higher than about 30% suggest an unacceptable level of non-isomorphism. Intermediate values suggest, but do not guarantee, that the derivative is worth pursuing. Normal probability plots can be helpful in this respect (Howell & Smith, 1992).

(4) Given a positive result from point (3), complete data may be recorded on the same or an equivalent derivative crystal. Again, it may be useful to record data to low resolution in the first instance. 4 Å resolution is again quite sufficient to solve the structure of a heavy-atom constellation using direct or Patterson methods, allowing the more complete characterization of the potential derivative.

(5) If the compound proves to be a useful derivative, data can then be recorded to higher resolution for the computation of phase information. It may not be appropriate to record data to the highest resolution as for the native protein. In this context, the strength of the data is of primary importance, and relatively weak data at high resolution may be less relevant.

Some practical points are highly relevant here. The ability to store and reuse vitrified crystals means that potential derivatives can first be screened at the lowest possible resolution, and the crystal can be preserved and used later only if the derivative proves to provide useful phase information. The final resolution for data collection will then depend on the degree of isomorphism. The wavelength, if tunable, should be set to a value just below the absorption edge in order to maximize the anomalous signal. The redundancy can play an important role, as it is useful to have a large number of independent measurements so that outliers in the native or derivative data can be excluded, as these can cause major problems in either the Patterson or direct-methods approaches for locating the heavy atom (Part 12).

9.1.13.2. Anomalous scattering, MAD and SAD

The requirements for collecting data with an intrinsically weak anomalous signal are several. The highest possible resolution should not be the primary consideration. The emphasis is on data quality, as it is necessary to measure very small differences in structure-factor amplitudes, which are already in themselves relatively weak. Important considerations include the following.

(1) Optimization of the wavelength, particularly for MAD experiments.
(2) Ensuring that the anomalous data are complete in terms of all possible Bijvoet pairs. This is not always addressed by the currently available data-processing software.
(3) High redundancy of measurements significantly enhances the quality of the signal, as this provides effective averaging of errors and allows the rejection of statistical outliers. The latter is especially important for direct-methods solution of the anomalous-scattering constellation.
(4) However, the crystal lifetime is finite owing to the effects of radiation damage, which can introduce changes in intensities of the same order as the anomalous signal. For SAD (single-wavelength anomalous dispersion)/MAD data, the exposures should be limited to ensure the data are complete before the onset of substantial damage. This may well mean that the resolution limit should be set more modestly than for native data.

For MAD experiments (Hendrickson, 1999; Smith, 1991), which can only be carried out at SR sites, the optimum number of wavelengths at which data should be recorded remains unclear. Given finite beam time, the trade-off may be between measuring with limited redundancy at several wavelengths as against higher redundancy at a smaller number of wavelengths, or even at one wavelength.

SAD represents the limiting case. All data are recorded at one wavelength, reducing the requirement for fine monochromatization and for fine tunability and stability. Now quality, especially in the form of redundancy, is the dominating factor since all phasing is based purely on a single anomalous difference for each reflection. In recent years, SAD phasing has come to predominate in the number of novel structures deposited in the PDB.

9.1.13.3. Molecular replacement

For the initial data required for molecular replacement (MR), high resolution is not essential. Firstly, the method depends on homologous models that are usually only an imperfect representation of the structure under investigation, hence high-resolution data cannot be accurately modelled and will only introduce noise into the analysis. Secondly, the rotation function, the first step in MR, is based on the representation of the Patterson function in terms of spherical harmonics, which is limited in its accuracy.

In contrast, it is vital for MR applications that the most intense low-resolution terms are measured. The lack of such reflections strongly affects the rotation- and translation-function computations, as the functions are based on Patterson syntheses involving the square of the structure-factor amplitudes, and are dominated by the largest terms. Elimination of the strongest few per cent of the low-resolution data may well prevent a successful solution by MR.

However, for refinement of structures solved by MR, it is important that data be recorded to a resolution sufficient to allow escape from the phase bias introduced by the model. This is a key point. There are many examples where collection of data to a higher resolution has enabled the refinement of an MR solution which would not refine at the lower resolution.

9.1.13.4. Definitive data for refinement of protein models

All structures benefit from the highest accuracy in their atomic coordinates to shed light on the details of their biological function. These may include substrate or inhibitor complexes and mutants as well as native proteins where the analysis requires the full potential of X-ray crystallography. Many of these crystals will not diffract to atomic resolution; nevertheless, all steps in a detailed crystal structure analysis are made simpler as the resolution and quality of the data are increased. This includes solution of the phase problem, interpretation of the electron-density maps and refinement of the model.

The most appropriate strategy for data collection involves decisions based on a complex and mutually dependent set of parameters including:

(1) Crystal quality and availability. If only one crystal is available, the choices are limited. If many are available, then some experimentation is recommended to select a high-quality sample. This is greatly aided by the recent introduction of automated sample changers and strategy software on a number of beamlines.
(2) Cryogenic vitrification. In many cases, this allows collection of data from a single crystal. If appropriate cryogenic conditions cannot be established, making it necessary to record room-temperature data, this can affect strategy dramatically, in that several crystals might well be required to record the target resolution and completeness.
(3) X-ray source and detector. The availability of these again places restrictions on the experiments that are tractable. An SR source will always provide better data, but has logistical problems of availability and access. For some problems, SR becomes *sine qua non* and a rotating anode is just insufficient. These include the use of MAD techniques, very small crystals, large and complex structures with large unit cells such as viruses, and where atomic resolution data are needed.
(4) Overall data-collection time allocated. This has an obvious overlap with point (3). In particular, if SR is to be used later, then the resolution limit on the home source may be modest.

If SR is not likely to be employed, then a higher resolution may be aimed for, requiring more time, and again dependent on the pressure on local resources.

Whatever the resource, it is good to define a strategy that will provide high completeness of the unique amplitudes at the highest resolution, with the realization that there may be some conflict between these two requirements owing to radiation damage.

9.1.13.5. A series of mutant or complex structures

The detailed geometry of the molecule is already known and the rather general effects of ligand binding or mutation can be initially identified at a relatively modest resolution and completeness. As with heavy-atom screening, it is often advisable to check that the desired complex or structural modification has been achieved by first recording data at low resolution.

However, if the analysis then proves to be of real chemical interest, with a need for accurate definition of structural features, the data should be subsequently extended in resolution and quality. As with the identification of isomorphous derivatives, this approach has benefited greatly from cryogenic vitrification, where the sample can be screened at low resolution and then preserved for subsequent use.

9.1.13.6. Atomic resolution applications

As for MAD data, the needs for atomic resolution data are extreme, but rather different in nature. Atomic resolution refinement is addressed in Chapter 18.4. Suffice it to say that by atomic resolution it is meant that meaningful experimental data extend close to 1 Å resolution. There are two principal reasons for recording such data. Firstly, they allow the refinement of a full anisotropic atomic model, leading to a more complete description of subtle structural features. Secondly, direct methods of phasing are dependent upon the principle of atomicity.

The problems to be faced include:

(1) The high contrast in intensities between the low- and high-angle reflections. This may be much larger than the dynamic range of the detector. If exposure times are long enough to give good counting statistics at high resolution, then the low-resolution spots will be saturated. The solution is to use more than one pass with different effective exposure times.

(2) The overall exposure time is often considerable and substantial radiation damage may finally result. The completeness of the low-resolution data is crucial, and it is strongly recommended to collect the low-resolution pass first as the time taken for this is relatively small.

(3) The close spacing between adjacent spots within the lunes on the detector, dependent on the cell dimensions. The only aid is to use fine collimation.

(4) The overlap of adjacent lunes at high diffraction angle, especially if a long cell axis lies along the beam direction. Using an alternative mount of the crystal is the simplest solution. Otherwise, the rotation range per image must be reduced, increasing the number of exposures. This was a problem with slow read-out detectors, but is largely alleviated with CCDs.

(5) For direct-methods applications, a liberal judgement of resolution limit should be adopted. Even a small percentage of meaningful reflections in the outer shells can assist the phasing. These weak shells can be rejected or given appropriate low weights in the refinement. The strong, low-resolution terms are vital for direct methods.

9.1.14. The importance of low-resolution data

The low-resolution terms define the overall shape of the object irradiated in the diffraction experiment. Omission of the low-resolution reflections, especially those with high amplitude, considerably degrades the contrast between the major features of the object and its surroundings. For a macromolecule, this means that the contrast between it and the envelope of the disordered aqueous solvent is diminished and, furthermore, the continuity of structural features along the polymeric chain may be lost. Refinement and analysis of macromolecules at all resolutions, be they high or low, involves the inspection of electron-density syntheses. These can be interpreted visually, on a graphics station, or automatically with a variety of software. In all of these, at all resolutions, the importance of the low-resolution terms is crucial. A special problem is in the interpretation of the partially ordered solvent interface. The biological activity of most enzymes and ligand-binding proteins is located precisely at this interface, and for a true structural understanding of how they function this region should be optimally defined. This is seriously impaired by the absence of the strong, low-resolution terms. The problems become more severe as the upper resolution limit of the analysis becomes poorer. Thus, at 1 Å resolution the omission of the 7 Å data shell will have less effect compared with a 3 Å analysis – but remember that ideally, no low-resolution data should be omitted!

In some phasing procedures, the presence of complete, especially high-intensity, low-resolution, data is even more crucial. The big, low-resolution amplitudes dominate the Patterson function, and methods based on the Patterson function are therefore especially sensitive. This encompasses one of the major techniques of phase determination for macromolecules: molecular replacement. Direct methods of phase determination utilize normalized structure factors and predominantly exploit those of high amplitude. The relations between the phases of those reflections with high amplitudes, such as the classical triple-product relationship, are strongest and most abundant for reflections with low Miller indices, hence at low resolution.

The importance of the low-resolution reflections in terms of geometric and qualitative context cannot be overemphasized.

9.1.15. Data quality over the whole resolution range

It is not possible to judge data quality from a single global parameter, especially $R_{\rm merge}$, not even from the overall $I/\sigma(I)$ ratio. Such a parameter may totally neglect problems such as the omission of all low-resolution terms due to detector saturation. A set of key parameters including $I/\sigma(I)$, $R_{\rm merge}$, percentage completeness, redundancy of measurements and number of overloaded high-intensity measurements must be tabulated in a series of resolution shells. This information should be assessed during data collection to guide the experimenter in the optimization of the choice of such parameters as exposure time, attainable resolution and required redundancy. As stated in Section 9.1.13, the requirements will vary with the application.

The effect of sample decay also requires such tables. The X-ray intensities decay more rapidly at high angle than at low, and consideration of this effect requires knowledge of the relative B values that need to be applied to the individual images during data scaling. An often subjective decision will need to be made regarding at what stage the decay is sufficiently high that further images should be ignored. The effects of damage are likely to be systematic rather than just random, and cannot be totally compensated for by scaling. This remains true even for cryogenically vitrified crystals, especially with ultra-bright synchrotron sources.

Following an earlier recommendation by the IUCr Commission on Biological Molecules (Baker *et al.*, 1996), this tabulated information, as a function of resolution, should be deposited with the data and the final model coordinates in the Protein Data Bank. Only then is it possible to have a true record of the experiment and for users of the database to judge the correctness and information content of a structural analysis.

9.1.16. Strategies for automated data acquisition

Progress in crystal-handling hardware has resulted in the development of sample changers both for synchrotron beamlines and the home laboratory. A sequence of samples contained in a Dewar can be mounted on the goniostat, centred in the beam and exposed to X-rays without user intervention. The gain in efficiency arises from the fact that manual intervention is no longer needed between samples, and at SR sites access to the hutch is avoided.

While this provides the potential for greatly increased throughput, it still requires intelligent decision-making software for evaluation of crystals and for optimal strategies of data collection to be achieved with minimal (ideally zero) user input. The steps required in such a system are:

(1) Automated mounting and dismounting of samples from a Dewar.
(2) Automated centring of the sample (or at least loop) in the beam.
(3) Recording and interpretation of two images preferably 90° apart in crystal orientation.
(4) Repetition of steps (1)–(3) for a number of samples of the same protein and ranking of these samples in terms of diffraction quality.
(5) Selection of the best sample and definition of the optimum strategy for data collection, taking account of information provided by the user with regard to the minimum resolution *etc*.
(6) Collection and integration of complete and ideally redundant data.

Robotic sample changers with at least some elements of the above are now operational at many sites [for example, see Leslie *et al.* (2002); McPhillips *et al.* (2002) and Cipriani *et al.* (2006), and also Chapter 9.2]. Considerable advances are expected in the near future, allowing routine automated screening of samples at major synchrotrons. In addition, remote access to synchrotons, by submission of crystals in Dewars for collection by SR staff (the so-called Fedex procedures) or for direct control by the user from their home laboratory through the Internet, is now possible at a number of synchrotron facilities. All of these moves towards automation require electronic databases for the tracking and transfer of samples and their associated data and parameters.

9.1.17. Final remarks

Optimal strategies for data collection are dependent on a number of factors. The alternative data-collection facilities to which access is potentially available, how long it takes to gain access and the overall time allocated all place restraints on the planning of the experiment. In view of this, it is not possible to provide absolute rules for optimal strategies.

Even after the source and overall time have been allocated or planned, the strategy is still the result of a compromise between several competing requirements. Some are general, others depend on the characteristics of a particular crystal or detector. As seen in the previous section, it is not possible to define protocols relevant for all applications. Rather, it is important to consider the relative importance of the parameters that can be varied to the problem in question and make the appropriate decisions.

Thus, data collection may have become easier from a technical point of view, but several crucial scientific decisions still have to be made by the experimenter. It is always beneficial to sacrifice some beam time and interpret the initial diffraction images, so as to avoid mistakes which may have an adverse effect on data quality and the whole of the subsequent structural analysis.

References

Amemiya, Y. (1995). *Imaging plates for use with synchrotron radiation. J. Synchrotron Rad.* **2**, 13–21.

Arndt, U. W. & Wonacott, A. J. (1977). Editors. *The Rotation Method in Crystallography.* Amsterdam: North Holland.

Baker, E. N., Blundell, T. L., Vijayan, M., Dodson, E., Dodson, G., Gilliland, G. L. & Sussman, J. L. (1996). *Deposition of macromolecular data. Acta Cryst.* D**52**, 609.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *On the conformation of the hen egg-white lysozyme molecule. Proc. R. Soc. London Ser. B*, **167**, 365–377.

Bloomer, A. C. & Arndt, U. W. (1999). *Experiences and expectations of a novel X-ray microsource with focusing mirror. I. Acta Cryst.* D**55**, 1672–1680.

Bourenkov, G. P. & Popov, A. N. (2006). *A quantitative approach to data-collection strategies. Acta Cryst.* D**62**, 58–64.

Broennimann, Ch., Eikenberry, E. F., Henrich, B., Horisberger, R., Huelsen, G., Pohl, E., Schmitt, B., Schulze-Briese, C., Suzuki, M., Tomizaki, T., Toyokawa, H. & Wagner, A. (2006). *The PILATUS 1M detector. J. Synchrotron Rad.* **13**, 120–130.

Carter, C. W. Jr & Sweet, R. M. (1997). Editors. *Methods in Enzymology*, Vol. 276, pp. 183–358. San Diego: Academic Press.

Cipriani, F., Felisaz, F., Launer, L., Aksoy, J.-S., Caserotto, H., Cusack, S., Dallery, M., di-Chiaro, F., Guijarro, M., Huet, J., Larsen, S., Lentini, M., McCarthy, J., McSweeney, S., Ravelli, R., Renier, M., Taffut, C., Thompson, A., Leonard, G. A. & Walsh, M. A. (2006). *Automation of sample mounting for macromolecular crystallography. Acta Cryst.* D**62**, 1251–1259.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

Cruickshank, D. W. J. (1999*a*). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (1999*b*). *Remarks about protein structure precision. Erratum. Acta Cryst.* D**55**, 1108.

Dauter, Z. (1999). *Data-collection strategies. Acta Cryst.* D**55**, 1703–1717.

Dauter, Z. (2005). *Efficient use of synchrotron radiation for macromolecular diffraction data collection. Prog. Biophys. Mol. Biol.* **89**, 153–172.

Diederichs, K. & Karplus, P. A. (1997). *Improved R-factor for diffraction data analysis in macromolecular crystallography. Nat. Struct. Biol.* **4**, 269–275.

Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Zero-dose extrapolation as part of macromolecular synchrotron data reduction. Acta Cryst.* D**59**, 903–909.

Evans, G. & Walsh, M. (2005). Editors. *Data Collection and Analysis. Proceedings of the CCP4 Study Weekend. Acta Cryst.* D**62**, 1–124.

Garman, E. F. & Owen, R. L. (2006). *Cryocooling and radiation damage in macromolecular crystallography. Acta Cryst.* D**62**, 32–47.

Garman, E. F. & Schneider, T. R. (1997). *Macromolecular cryocrystallography. J. Appl. Cryst.* **30**, 211–237.

Gruner, S. M. & Ealick, S. E. (1995). *Charge coupled device X-ray detectors for macromolecular crystallography. Structure*, **3**, 13–15.

Hamlin, R. (1985). *Multiwire area X-ray diffractometers. Methods Enzymol.* **114**, 416–452.

Helliwell, J. R. (2004). *Macromolecular Crystallography with Synchrotron Radiation.* Cambridge University Press.

Henderson, R. (1990). *Cryo protection of protein crystals against radiation damage in electron and X-ray diffraction. Proc. R. Soc. London Ser. B*, **241**, 6–8.

Hendrickson, W. A. (1999). *Maturation of MAD phasing for the determination of macromolecular structures*. *J. Synchrotron Rad.* **6**, 845–851.

Hendrickson, W. A. (2000). *Synchrotron crystallography*. *Trends Biochem. Sci.* **25**, 637–643.

Howell, P. L. & Smith, G. D. (1992). *Identification of heavy-atom derivatives by normal probability methods. J. Appl. Cryst.* **25**, 81–86.

Hülsen, G., Broennimann, C., Eikenberry, E. F. & Wagner, A. (2006). *Protein crystallography with a novel large-area pixel detector*. *J. Appl. Cryst.* **39**, 550–557.

International Tables for Crystallography (2005). Vol. A. *Space-Group Symmetry*, edited by Th. Hahn. Heidelberg: Springer.

Kabsch, W. (1988). *Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. J. Appl. Cryst.* **21**, 916–924.

Leslie, A. G. W., Powell, H. R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. & Nave, C. (2002). *Automation of the collection and processing of X-ray diffraction data – a generic approach*. *Acta Cryst.* D**58**, 1924–1928.

McPhillips, T. M., McPhillips, S. E., Chiu, H.-J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M. & Kuhn, P. (2002). *Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines*. *J. Synchrotron Rad.* **9**, 401–406.

Matthews, B. W. (1968). *Solvent content in protein crystals. J. Mol. Biol.* **33**, 491–497.

Popov, A. N. & Bourenkov, G. P. (2003). *Choice of data-collection parameters based on statistic modelling*. *Acta Cryst.* D**59**, 1145–1153.

Read, R. J. (1999). *Detecting outliers in non-redundant diffraction data. Acta Cryst.* D**55**, 1759–1764.

Rodgers, D. W. (1997). *Practical cryocrystallography. Methods Enzymol.* **276**, 183–203.

Rosenbaum, G., Holmes, K. C. & Witz, J. (1971). *Synchrotron radiation as a source for X-ray diffraction. Nature (London)*, **230**, 434–437.

Schiltz, M., Dumas, P., Ennifar, E., Flensburg, C., Paciorek, W., Vonrhein, C. & Bricogne, G. (2004). *Phasing in the presence of severe site-specific radiation damage through dose-dependent modelling of heavy atoms*. *Acta Cryst.* D**60**, 1024–1031.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Smith, J. L. (1991). *Determination of three-dimensional structure by multiwavelength anomalous diffraction. Curr. Opin. Struct. Biol.* **1**, 1002–1011.

Weiss, M. S. (2001). *Global indicators of X-ray data quality*. *J. Appl. Cryst.* **34**, 130–135.

International Tables for Crystallography (2012). Vol. F, Chapter 9.2, pp. 231–233.

Chapter 9.2. Robotic crystal loading

T. Earnest and C. Cork

9.2.1. Introduction

The three-dimensional structure of a biological molecule provides fundamental information about how it interacts with other molecules, co-factors and ligands that defines, at a molecular level, how the protein functions in the cellular environment. Protein structure information has also proved to be an invaluable tool in the development of new synthetically derived medicines. The determination of protein structures has significantly accelerated in the past decade, due in large part to advances in protein production (cloning, sequencing, expression and purification), protein crystallography (*e.g.* cryogenic freezing, intense synchrotron X-ray sources, area detectors) and computing. There has been an explosive growth in the use of synchrotron sources for the collection of X-ray diffraction intensities from protein crystals, since synchrotrons can provide bright tunable X-rays which are needed for both high-throughput projects and challenging biomolecular targets such as large complexes and membrane proteins (Wakatsuki & Earnest, 2000).

The convergence of a number of factors in the 1990s and early 2000s motivated the development of robotic crystal mounting – the increase in the number and performance of synchrotron beamlines for biological crystallography, improvements in detector speed, the motorization and computer control of beamlines, advances in computer hardware and software, and an increase in the demand for beam time as structural biologists began to pursue ever more difficult crystallographic projects. Structural genomics efforts and structure-based drug-design programmes benefited significantly, with an increase in the throughput of data collection and analysis for which these systems provided an enabling technology. During the 1990s, the use of cryogenic data collection, instead of mounting in glass capillaries, also allowed synchrotron beamlines to be used more productively, since the bright beams lead very rapidly to the onset of radiation damage unless the crystal is preserved and maintained at low temperatures, typically ~100–120 K.

9.2.2. Robotic sample loaders

Significant amounts of synchrotron beam time can be lost due to crystal manipulation and misalignment by the beamline user, which becomes more critical with the increase in brightness and stability of the newer synchrotron sources. Crystal-screening and data-collection runs have become significantly shorter. Therefore, more efficient mounting and alignment tools have been developed that reduce the fraction of time spent changing samples, including the implementation of robotic sample loaders (automounters), which can greatly facilitate the throughput and production of crystallography beamlines and other X-ray sources.

There are several benefits to using an automounter system, particularly on bright synchrotron beamlines:

(i) *It facilitates optimum use of synchrotron beam time.* Crystallography experiments are installed in radiation-shielded hutches, which are inaccessible during data collection. Changing the crystal manually requires opening and closing the hutch, initiating the interlocks and performing a hutch search, which typically takes several minutes. The automounter significantly reduces both the sample mounting time and the number of required hutch accesses.

(ii) *It facilitates advanced data-collection techniques.* An experimental station can be fully automated, including integrated data collection and processing whereby the structure-solving software can influence the data-collection process (*e.g.* crystal ranking and data-collection strategy determination). Remote data collection, where the researchers send cryoprotected crystals (maintained near liquid-nitrogen temperatures) to the beamline for robotic mounting and automated (or semi-automated) data collection, can proceed in a flexibly scheduled manner.

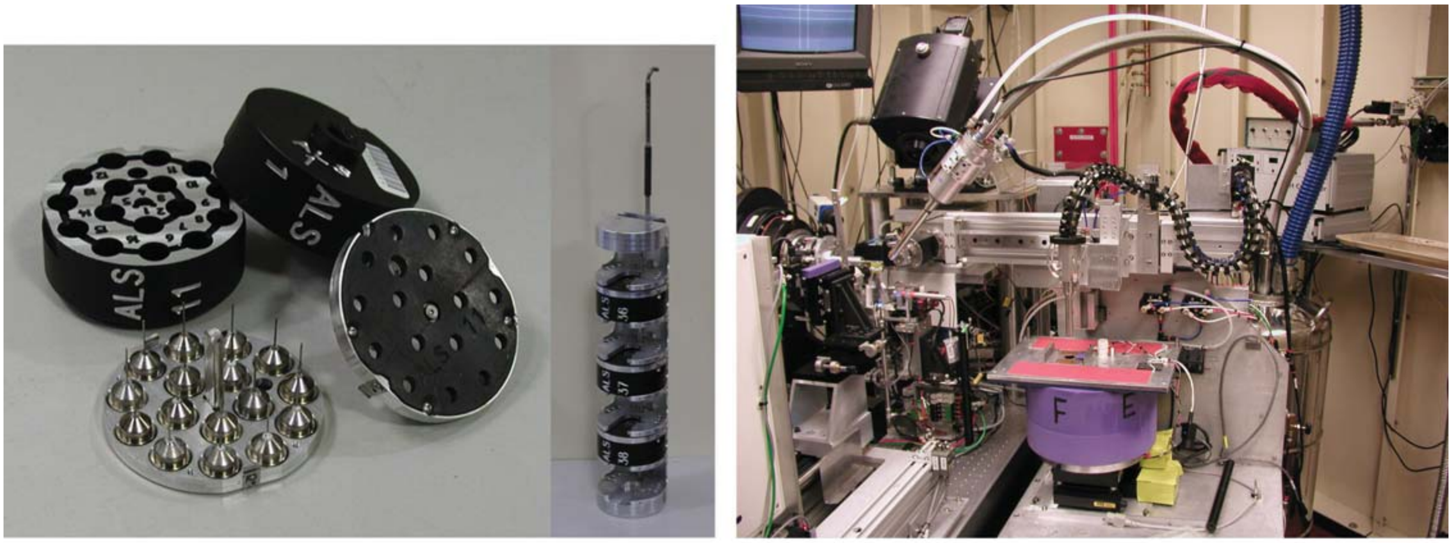
(iii) *It facilitates the collection of higher-quality data.* The rapid crystal-interchange mechanism enables the researcher to evaluate a large pool of samples and to select the best crystals from the set.

(iv) *It reduces risk to crystals.* Automated mounting and dismounting of crystals can be done much more reliably than manual handling.

(v) *It facilitates systematic studies of experimental protocols.* Alternative protocols can be performed in a manner and number that would be impractical for humans to perform manually. Furthermore, this can allow for an intelligent system to 'learn' improved methods of data collection and processing.

The major challenge in the design of automation hardware for the mounting and alignment of crystals of biological samples is the necessity to maintain the sample at approximately 100 K, since cryogenic data collection is required at synchrotron beamlines and usually at home sources as well. The process involves the optical screening of several candidate crystals under a microscope, followed by 'freezing' the selected crystal in an amorphous glass formed by the mixing of reservoir and cryoprotectant solutions. In this manner, it is possible to manipulate small or fragile crystals more easily. More importantly, the frozen crystals are far more resistant to X-ray induced radiation damage, a common problem at synchrotron sources. The frozen crystals can then be stored indefinitely in liquid nitrogen and handled with minimum difficulty at cryogenic temperatures. There are established procedures for the manual manipulation and mounting of the crystal while at cryogenic temperatures. At the beamline, the crystals are transferred from the transport container to a small holding Dewar. Next, they are mounted in a secondary sample-transfer tool, and then they are transferred to the X-ray diffractometer. These can be very time-consuming steps when performed manually. Consequently, instrumentation has been developed to minimize the extent of manual manipulations of frozen crystals at the beamline (Snell *et al.*, 2004; Cork *et al.*, 2006) (Fig. 9.2.2.1). Groups at Abbott (Muchmore *et al.*, 2000), at the SSRL (Cohen *et al.*, 2002), at European facilities (Cipriani *et al.*, 2006) and in Japan (Ueno *et al.*, 2006) have similarly produced robotic crystal-mounting systems with a diversity of approaches to achieve the same goal. Commercial

Copyright © 2012 International Union of Crystallography



(*a*) (*b*)

Figure 9.2.2.1
(*a*) Hardware for the storage and transport of crystals. Each puck holds up to 16 crystals, and seven can be shipped in a standard cryo-shipper. (*b*) The robotic crystal mounter implemented on Beamline 5.0.2 at the Advanced Light Source.

systems have also been developed (*e.g.* the Rigaku ACTOR and the IRELEC CATS).

There are two basic approaches that have been pursued: (i) the use of commercial multi-axis robots (Muchmore *et al.*, 2000; Cohen *et al.*, 2002; Rigaku ACTOR) and (ii) modular purpose-built systems, such as the one developed at the Lawrence Berkeley Laboratory Advanced Light Source and shown in Fig. 9.2.2.1(*b*) (Snell *et al.*, 2004). All systems must take into consideration that the procedures used to transfer samples from a crystal tray and mount them on the goniometer need to ensure that the crystals be maintained in a frozen frost-free environment and transferred rapidly. For example, the system developed at the Advanced Light Source can be adapted for use with a variety of experimental station configurations. The only significant constraint in the use of this design is that the nitrogen cold stream subsystem must be oriented off-axis from the primary goniometer axis. The first automounter was installed on experimental station BL5.0.3 at the Advanced Light Source, a fixed wavelength station with significant access constraints. This system has been in operation since March 2001. Shortly thereafter, systems were installed on a second monochromatic experimental station, BL5.0.1, and also on a multiwavelength experimental station, BL5.0.2 (Snell *et al.*, 2004). The automounter configuration is essentially identical for all three experimental stations with a high degree of standardization. The sample-mounting robot consists of three main components: the gripper, which holds the sample during transport from the Dewar to the goniometer while keeping it at low temperature and frost-free; the X–Y–θ stage, which performs the actual transport; and the Dewar stage, which supports the samples in a regular array submerged in liquid nitrogen and positions the selected sample for access by the transporter (Fig. 9.2.2.1*b*).

The gripper contains a conically shaped brass collet that can be opened and closed using a small pneumatic actuator (Fig. 9.2.2.2). The collet is precooled to ~100 K before engaging the sample holder. A sensor inside the collet is used to monitor its temperature. The inner tube is surrounded by a low thermal capacity outer shroud (a very thin stainless steel tube), which provides a sheath flow of warm dry gas to reduce icing and frost formation during exposure to ambient moisture-laden air. During continuous use, ice is periodically removed from the gripper by means of an air-jet heater (*cf.* Fig 9.2.2.1*b*). Additionally, the gripper has a small low-force linear stage (labelled 'small move' in Fig. 9.2.2.2). This stage is used for the final positioning of the gripper on the sample during mounting and dismounting. The sample gripper was originally designed around the standard Hampton metal cap, which is held on the goniometer by a permanent magnet. A modified cap design, which has tighter tolerances and a conical shape rather than a ledge, has been

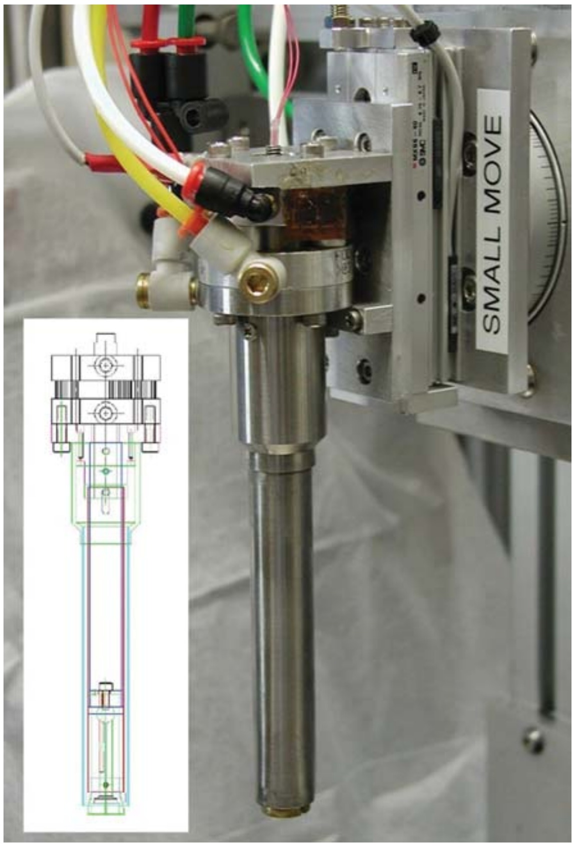


Figure 9.2.2.2
Crystal gripper assembly with low-force actuator (labelled 'small move').

designed to increase automounter reliability (also available from Hampton Research as part No. HR4-779).

The gripper is mounted on a pneumatic X–Y–θ stage, which is used to transport samples between the goniometer and the storage Dewar. A vertical Y stage moves the gripper in and out of the Dewar, a 90° rotational θ stage orients the gripper either horizontally or vertically, and a long horizontal X stage moves the gripper between the Dewar and the goniometer mount points.

When they are not mounted on the goniometer for data collection, the samples are maintained in a small cylindrical Dewar (see Fig. 9.2.2.1*b*). A custom cassette or puck (Fig. 9.2.2.1*a*) facilitates automated handling and bulk transport. Up to seven cassettes (112 samples) can be loaded into the sample Dewar. The Dewar is mounted on an R–θ motorized stage, which is used to position the selected sample for access by the gripper. The Dewar can also be positioned such that the gripper will be inserted into an unoccupied space for a precooling action when boiling of liquid nitrogen might endanger the samples. The Dewar is automatically filled directly from the facility's liquid-nitrogen supply system. During normal operations, an insulated and heated cover reduces icing and evaporation of liquid nitrogen. The gripper reaches into the Dewar through a small heated hole in the cover. The sample cassettes are moved in or out of the Dewar through a small access port in the cover.

The Robosync web site (http://smb.slac.stanford.edu/robosync/), maintained by the SSRL Structural Molecular Biology Group, features a list of biological crystallography beamlines with automated crystal mounting systems.

Important for the efficient and productive use of robotic crystal mounters is their integration into the beamline control system (McPhillips *et al.*, 2002; Snell *et al.*, 2004). The different motions of the automounter can typically be actuated from a software interface, either individually or through scripts which run a sequence of actions. The development of more fully automated capabilities requires the integration of the automounter and beamline control system with the data collection and analysis software. To this end, software has been developed that provides a framework for indexing integration and symmetry determination, along with suggested strategies for data collection (González *et al.*, 2008). When coupled with bright synchrotron beamlines and automounters, the underlying requirements for building fully automated and optimized data collection are mostly in place. The development of intelligent automated control (through *e.g.* artificial-intelligence approaches) should permit fully automated data-collection facilities to collect optimized data from a wide range of structural projects in the most efficient and cost-effective way and with minimal human intervention prior to the scientific analysis of the structures obtained.

9.2.3. Conclusion

In summary, the development of automounters over the past decade has been driven by the demands of greater access to synchrotron resources, the need for high-throughput data collection from projects that require numerous structures (*e.g.* structure-based drug-design and structural genomics programmes) and the demonstrated improvement in data quality for more challenging projects (such as large complexes and membrane proteins) that require the screening of numerous crystals in order to select those with the best diffraction characteristics. A number of different designs have been implemented at synchrotron sources worldwide, to the significant advantage of the structural biology community.

The authors thank their collaborators at the LBNL Engineering Division, the Berkeley Center for Structural Biology, the Advanced Photon Source, the National Synchrotron Light Source and the Cornell High Energy Synchrotron Source, as well as the numerous users who provided valuable testing and comments. We also thank the National Institute of General Medical Sciences of the National Institutes of Health for generous long-term support for this scientific engineering project.

References

Cipriani, F., Felisaz, F., Launer, L., Aksoy, J.-S., Caserotto, H., Cusack, S., Dallery, M., di-Chiaro, F., Guijarro, M., Huet, J., Larsen, S., Lentini, M., McCarthy, J., McSweeney, S., Ravelli, R., Renier, M., Taffut, C., Thompson, A., Leonard, G. A. & Walsh, M. A. (2006). *Automation of sample mounting for macromolecular crystallography. Acta Cryst.* D**62**, 1251–1259.

Cohen, A. E., Ellis, P. J., Miller, M. D., Deacon, A. M. & Phizackerley, R. P. (2002). *An automated system to mount cryo-cooled protein crystals on a synchrotron beamline, using compact sample cassettes and a small-scale robot. J. Appl. Cryst.* **35**, 720–726.

Cork, C., O'Neill, J., Taylor, J. & Earnest, T. (2006). *Advanced beamline automation for biological crystallography experiments. Acta Cryst.* D**62**, 852–858.

González, A., Moorhead, P., McPhillips, S. E., Song, J., Sharp, K., Taylor, J. R., Adams, P. D., Sauter, N. K. & Soltis, S. M. (2008). *Web-Ice: integrated data collection and analysis for macromolecular crystallography. J. Appl. Cryst.* **41**, 176–184.

McPhillips, T. M., McPhillips, S. E., Chiu, H.-J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M. & Kuhn, P. (2002). *Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. J. Synchrotron Rad.* **9**, 401–406.

Muchmore, S. W., Olson, J., Jones, R., Pan, J., Blum, M., Greer, J., Merrick, S. M., Magdalinos, P. & Nienaber, V. L. (2000). *Automated crystal mounting and data collection for protein crystallography. Structure*, **8**, R243–R246.

Snell, G., Nordmeyer, R., Cornell, E., Meigs, G., Cork, C., Yegian, D., Jaklevic, J., Jin, J., Stevens, R. C. & Earnest, T. (2004). *Automatic sample mounting and alignment system for biological crystallography at a synchrotron source. Structure*, **12**, 536–545.

Ueno, G., Kanda, H., Hirose, R., Ida, K., Kumasaka, T. & Yamamoto, M. (2006). *RIKEN structural genomics beamlines at the SPring-8; high throughput protein crystallography with automated beamline operation. J. Struct. Funct. Genom.* **7**, 15–22.

Wakatsuki, S. & Earnest, T. N. (2000). *The impact of synchrotron radiation in protein crystallography. Synchrotron Radiat. News*, **13**, 4–8.

International Tables for Crystallography (2012). Vol. F, Chapter 9.3, pp. 234–239.

Chapter 9.3. X-ray diffraction imaging of whole cells

D. Shapiro

9.3.1. Introduction

The brightness of third-generation synchrotron X-ray sources removes the need to enhance a diffracted X-ray signal with crystallographic redundancy for certain classes of samples. This is particularly useful for samples for which there exists only one unique structure or for which arrangement into a crystalline form is exceedingly difficult. Single-particle X-ray diffraction microscopy, also known as coherent X-ray diffraction microscopy (CXDM), treats an isolated non-crystalline sample as a crystallographer treats a crystal. The far-field diffraction intensity pattern of the sample is measured and the phase problem is solved computationally, allowing for structure recovery through a Fourier transform. Since Mother Nature restricts us to only measuring intensities, all methods of phase retrieval rely on decoding phase information that has been encoded in intensity measurements. Crystallographers, and the subsequent diffraction microscopists, have two basic methods for doing this. The first, the holographic method, mixes a known reference signal with the desired and unknown signal. The coherent superposition of these two signals results in measureable intensity variations that are directly related to the relative phase. This is, indirectly, a phase-measuring method. The second method relies only on the signal from the unknown structure and additional information, supplied by the scientist, which is physically plausible. This is a constraint-based method. Both methods seek to fill the information deficit inherent in intensity measurements by adding something that may unlock the phase.

In 1980, David Sayre suggested that CXDM should be possible (Sayre, 1980). Algorithms being developed for electron microscopy in the early 1970s would establish the constraint-based paradigm of iterative phase retrieval from Fourier modulus measurements. In particular, the alternating projection algorithm of Gerchberg and Saxton, the error-reduction algorithm, was developed to reconstruct the phase information that is missing when intensity measurements are made in both real and reciprocal space (Gerchberg & Saxton, 1972). This algorithm was later modified by Fienup into the input–output algorithm to handle cases where only one intensity measurement is made (Fienup, 1978, 1982; Fienup *et al.*, 1982; Miao *et al.*, 1998). In this case, the required real-space constraint restricts the object to an area no larger than half the width of its autocorrelation. The combination of the input–output algorithm and the error-reduction algorithm was found to be a very robust method of image reconstruction using only Fourier domain intensities, but its success was not understood for several years. In 1982, Bates argued that the solutions to the phase problem are unique in two dimensions if the Fourier modulus is sampled on an interval at least twice as fine as the Bragg interval (Bates & Fright, 1983; Bates, 1982). The method of phase retrieval from diffraction patterns sampled between Bragg peaks became known as the oversampling phasing method (Sayre, 1991; Sayre & Chapman, 1995; Miao *et al.*, 1998). The oversampling method compensates for the information deficit by supplying a zero-density region of approximately known size in the object domain. The more finely the diffraction pattern is sampled, the larger this zero-density region is, although this only adds useful information up to a certain point. This addition of information through Fourier space sampling results in an overdetermined inverse problem with a unique solution.

Sayre considered that the real niche of CXDM would be imaging objects of a few microns at a resolution of a few nanometres using soft X-rays (1–10 nm wavelength). Indeed, since the initial demonstration of CXDM (Fig. 9.3.1.1), ten years ago at the time of writing, active research projects have developed at all major synchrotron facilities and the technique has been applied to a diverse set of scientific problems. Three-dimensional images of radiation-hard materials have been obtained at 15 nm resolution, while two-dimensional images of a whole cell at 30 nm and of a gold nanoparticle at 5 nm resolutions have been reported, among many others (Miao *et al.*, 2002; Shapiro *et al.*, 2005; Pfeifer *et al.*, 2006; Barty *et al.*, 2008; Schroer *et al.*, 2008; Nelson *et al.*, 2010). CXDM researchers have not yet achieved the 'holy grail' of cellular imaging, the imaging of a whole frozen hydrated cell in three dimensions, but at least two teams are pursuing this goal with recent success in the two-dimensional case (Huang *et al.*, 2009; Lima *et al.*, 2009). Lens-based X-ray microscopes have achieved this landmark with moderate resolution (Wang *et al.*, 2000; Weiss *et al.*, 2000; Larabell & Le Gros, 2004). Transmission X-ray microscopes (TXM) and scanning transmission X-ray microscopes (STXM) utilizing diffractive zone plate lenses have high throughput and the advantage of direct imaging, but are limited in resolution by the technological challenge of making efficient high-numerical-aperture lenses. TXM can now routinely image in three dimensions at 50 nm resolution, and 12 nm resolution has

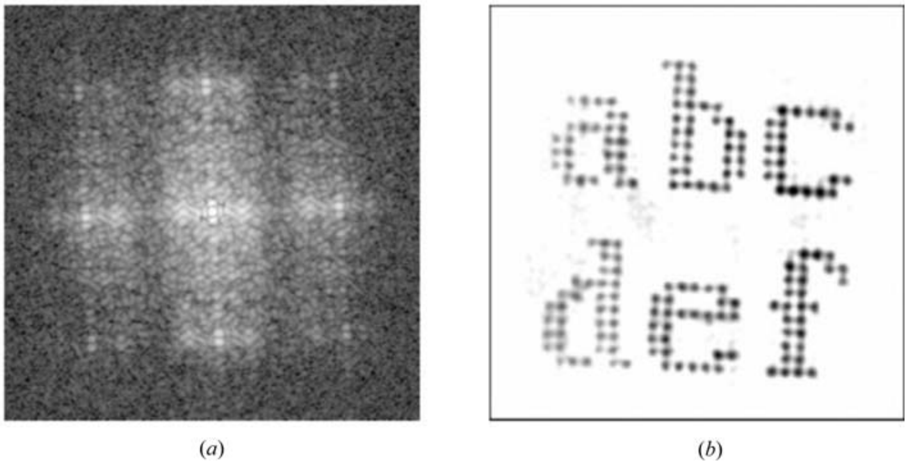


Figure 9.3.1.1
First soft X-ray demonstration of the CXDM method. (*a*) Diffraction pattern using 1.7 nm X-rays. (*b*) Reconstruction of (*a*) to 75 nm resolution. Reprinted by permission from Macmillan Publishers Ltd: *Nature* (Miao *et al.*, 1999), copyright (1999).

Copyright © 2012 International Union of Crystallography

been demonstrated in two dimensions, although the total efficiency of the optic used was of the order of 1% (Chao *et al.*, 2009) at the highest spatial frequencies. Thus, at the cost of throughput and ease of use, a diffraction microscope provides increased X-ray efficiency and resolution.

Table 9.3.2.1
Summary of various algorithms

The algorithms are, from top to bottom: error reduction, solvent flipping, hybrid input–output, difference map, averaged successive reflections, hybrid projection–reflection and relaxed averaged alternating reflections. A reflection is defined by its associated projection as $\mathbf{R} = 2\mathbf{P} - \mathbf{I}$, where $\mathbf{I}$ is the identity projection (Marchesini, 2007).

Algorithm	Iteration $\rho^{(n+1)} =$
ER	$\boldsymbol{P}_s\boldsymbol{P}_m\rho^{(n)}$
SF	$\boldsymbol{R}_s\boldsymbol{P}_m\rho^{(n)}$
HIO	$\begin{cases} \boldsymbol{P}_m\rho^{(n)}(\boldsymbol{r}), & \boldsymbol{r} \in S \\ (\boldsymbol{I} - \beta\boldsymbol{P}_m)\rho^{(n)}(\boldsymbol{r}), & \boldsymbol{r} \notin S \end{cases}$
DM	$\{\boldsymbol{I} + \beta\boldsymbol{P}_s[(1+\gamma_s)\boldsymbol{P}_m - \gamma_s\boldsymbol{I}]$ $-\beta\boldsymbol{P}_m[(1+\gamma_m)\boldsymbol{P}_s - \gamma_m\boldsymbol{I}]\}\rho^{(n)}$
ASR	$(1/2)[\boldsymbol{R}_s\boldsymbol{R}_m + \boldsymbol{I}]\rho^{(n)}$
HPR	$(1/2)\{\boldsymbol{R}_s[\boldsymbol{R}_m + (\beta - 1)\boldsymbol{P}_m] + \boldsymbol{I} + (1-\beta)\boldsymbol{P}_m\}\rho^{(n)}$
RAAR	$[(1/2)\beta(\boldsymbol{R}_s\boldsymbol{R}_m + \boldsymbol{I}) + (1-\beta)\boldsymbol{P}_m]\rho^{(n)}$

The following sections first report on the current standard single-particle phase retrieval techniques and then on recent experiments in CXDM, which establish the state of the art in whole-cell imaging by diffractive methods. Images of dry yeast at 11 nm resolution are presented, which represent the highest resolution X-ray images of whole cells currently on record. The effects of radiation damage are discussed and Sayre's idea of using stereoscopic viewing as a means of obtaining quick and low-dose three-dimensional information is explored (Sayre, 2008).

9.3.2. Phase retrieval from single-particle diffraction data

The problem of phase retrieval is solved through successive application of constraints on the recovered object in the data and object spaces. Using the language of convex optimization, the mathematical operators which act on the data are projectors. The projector in the reciprocal (data) space forces the Fourier components to have the correct magnitude, while in object space finite support is enforced. To calculate the Fourier magnitude projector, one first needs to propagate the object density, ρ, to the data space by a Fourier transform, then replace the estimated magnitudes $|\tilde{\rho}|$ with the measured ones, $I^{1/2}$, and finally propagate back to real space. Using these transforms one simplifies the calculation of the projection, which becomes an element-wise operation on each recovered Fourier component. The forward $\mathcal{F}$ and inverse $\mathcal{F}^{-1}$ transforms must be incorporated into the operator defined in real space $\mathbf{P}_m$,

$$\mathbf{P}_m = \mathcal{F}^{-1}\tilde{\mathbf{P}}_m\mathcal{F}, \qquad (9.3.2.1)$$

where the measured Fourier magnitudes are enforced in Fourier space by $\tilde{\mathbf{P}}_m$. Using the Fourier basis, one simply replaces the estimated magnitudes $|\tilde{\rho}|$ with the measured ones $I^{1/2}$, $[\tilde{\mathbf{P}}_m\tilde{\rho}(\boldsymbol{k}) = I(\boldsymbol{k})^{1/2}\tilde{\rho}(\boldsymbol{k})/|\tilde{\rho}(\boldsymbol{k})|]$. Similarly, in the object space the finite support constraint is applied on a per pixel basis through multiplication by the support mask. The corresponding projector is

$$\mathbf{P}_s\rho = S \cdot \rho.$$

Table 9.3.2.1 lists the combination of projections used by the most popular algorithms.

The violation of the support constraint is used as an error metric to monitor the convergence towards the solution. The solution should have zero density outside the support mask, so the error can be defined as the total density outside the support area,

$$\varepsilon_s^2(\rho) = |||\rho - S\rho|||^2 = ||[I - \mathbf{P}_s]\rho||^2. \qquad (9.3.2.2)$$

Alternatively, the error metric can be defined in the data space as the difference between the measured and calculated magnitudes,

$$\varepsilon_m^2(\rho) = |||\mathcal{F}\rho| - I^{1/2}||^2 = ||[I - \mathbf{P}_m]\rho||^2. \qquad (9.3.2.3)$$

In reality, the measured intensities are subject to noise which prohibits exact compliance with the constraints, so the error metrics cannot drop to zero. Although in most cases the algorithm can locate the global minimum, random noise will force fluctuations around the minimum. Once the algorithm reaches this steady-state regime, any particular iterate chosen as the solution would have a misleading degree of detail. On the other hand, the average of many fluctuating iterates would have reduced intensity in those Fourier components which are not reliably phased. The ratio of the average Fourier magnitude to the measured magnitude provides a measure of the stability, or reproducibility, of the retrieved phase information. This ratio as a function of spatial frequency is the phase retrieval transfer function (PRTF),

$$\mathrm{PRTF}(\mathbf{q}) = \frac{|\mathcal{F}[\langle\psi\rangle]|(\mathbf{q})}{[I_m(\mathbf{q})]^{1/2}}, \qquad (9.3.2.4)$$

where $\mathcal{F}[\langle\psi\rangle]$ is the Fourier transform of the final averaged image and I_m is the measured intensity pattern. The PRTF is analogous to the differential phase residual of electron microscopy and, following Chapman *et al.* (2006), the resolution of a reconstruction is chosen as the spatial frequency at which the PRTF falls below a value of 0.5.

9.3.3. High-resolution imaging of yeast

The X-ray dose required to image a given volume of protein is nearly independent of energy above the oxygen K edge. At the same time, the photon flux required to image the same volume increases with E^2 because of the energy dependence of the scattering cross section (Howells *et al.*, 2009). For this reason, it is advantageous to use the lowest energy commensurate with the desired resolution of 5–10 nm. Commercially available charge-coupled device (CCD) detectors can easily provide a scattering angle of 0.1 radians [a 1 inch detector placed 5 inches from the sample (1 inch = 2.54 cm)], which results in a half-period resolution of 8 nm when using 750 eV X-rays. Furthermore, a cell with a diameter of 3 µm would have an oversampling ratio (number of intensity samples per speckle) of at least ten in this geometry if the detector has 20 µm pixels.

In this particular case, a Princeton Instruments CCD (PI-MTE:1300) is placed 136 mm downstream of a freeze-dried yeast cell using the CXDM instrument on Beamline 9.0.1 of the Advanced Light Source (ALS). The yeast cell is illuminated by a coherent beam of 750 eV X-rays defined by a 5 µm pinhole located 25 mm upstream. The incident intensity of 4×10^6 photons s^{-1} µm^{-2} is high enough to cause rapid structural changes to the cell (discussed in the next section), so the sample requires pre-irradiation for about 30 minutes prior to collection of the final data set intended for reconstruction. The final data set, shown in Fig. 9.3.3.1, is a 1024 × 1024 pixel subset of the full CCD and extends to a resolution of 11 nm at the edge and 7.8 nm in the corner. Speckles extend to the corner of the data set after a total exposure of 406 s and to the edge after 226 s.

The full reconstruction of this type of data set can only proceed once a high-fidelity support is determined. Algorithmic support determination using *Shrinkwrap*, a variation on the hybrid input–output algorithm, is straightforward for objects with sharp boundaries, as is often the case in the material sciences (Marchesini *et al.*, 2003; Chapman *et al.*, 2006). However, biological samples which often have soft boundaries require manual intervention in the early stages of reconstruction. The soft-edge problem is exacerbated by the loss of low spatial frequency information behind the beamstop. A combination of *Shrinkwrap* support adjustments and intuitively reasonable manual adjustments are made until the algorithm is stable to automatic adjustment. This point is found when further *Shrinkwrap* adjustments no longer alter the shape of the support but just its tightness. High-fidelity reconstructions of complex-valued objects are not possible with a loose support (Fienup, 1987). Once the support is found, the final image is produced as the average over many reconstructions, all using the same support, which are started from different random phase sets. This averaging procedure, discussed earlier, reduces features that are primarily due to noise and provides a measure of the reproducibility of the recovered phases and therefore an estimate of the resolution. The final reconstruction shown in Fig. 9.3.3.1 had PRTF values above 0.5 for spatial frequencies extending to the edge of the data set or 11 nm resolution. The inset image clearly shows features of the order of 15 nm in size. In general, it is not possible to identify cell organelles without labelling specific proteins. However, a correlative study of yeast which combines high-resolution CXDM with optical fluorescence would be a powerful tool for the cell-biology community. This technique has already been demonstrated at lower resolutions using a transmission X-ray microscope (Le Gros *et al.*, 2009).

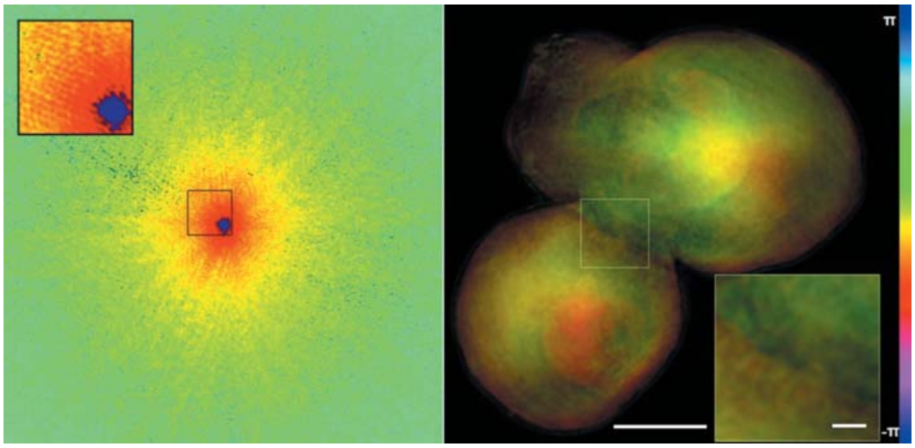


Figure 9.3.3.1
Diffraction pattern (left) and reconstruction (right) of a freeze-dried budding yeast cell. The diffraction pattern, measured on Beamline 9.0.1 of the ALS, extends to a half-period resolution of 11 nm and required 226 s of X-ray exposure. The blue regions of the diffraction pattern represent zeroes (noisy measurements or pixels lost behind the beamstop) and were left unconstrained during phase retrieval. The image represents the complex-valued X-ray wavefield after passing completely through the scattering potential, but propagated to the interior plane with the smallest support. The X-ray phase is represented as image hue and magnitude as brightness. The PRTF (not shown) never dips below 0.5, indicating that the magnitudes were adequately phased to the corner of the recorded data. The large scale bar is 1 µm, while the inset scale bar is 100 nm. The reconstruction presented here is the average of 25 independent reconstructions, each starting with a different set of random phases. Each reconstruction required 2000 iterations of the hybrid input–output algorithm and took about 14 s to complete on an nVidia Tesla C1060 graphics processing unit.

9.3.3.1. Radiation damage

High-resolution imaging of single particles with X-rays requires a large radiation dose because of the very strong dependence of the scattering cross section on spatial frequency and because, in the single-particle case, there is none of the coherent amplification one obtains when many identical copies of the particle are arranged into a crystal. This large dose means that either the sample must be protected from morphological changes induced by radiation exposure, most effectively through cryogenic techniques, or that the obtained images will represent an altered form of the original sample. Cryoprotection of hydrated cells has been successfully used in electron microscopy and lens-based X-ray microscopy for some time, but the sample preparation requirements of diffractive imaging are more severe, so the development of cryotechniques has been slower. The requirements of a finite sample support and that the sample be maintained on a zero-scattering background are extraordinarily difficult to achieve when that background consists of a micron-thick layer of ice. This is an active area of research by several X-ray diffraction microscopy groups.

For the case of dry cells, various preparation techniques (freeze drying or chemical fixation and dehydration) can preserve the large-scale internal structure, such as the size and shape of large organelles, but the ultrastructure will inevitably show artifacts of the drying process. Furthermore, exposure to ionizing radiation results in the well known shrinkage problem. Studies with a transmission X-ray microscope indicate that X-ray induced shrinkage primarily produces a higher-density but smaller version of the original cell (Jearanaikoon & Abraham-Peskir, 2005). The effect of cell shrinkage on the X-ray diffraction pattern is shown in Fig. 9.3.3.2. A freeze-dried cell is repeatedly exposed to 750 eV X-rays from Beamline 9.0.1 of the ALS. Each exposure is 30 s and delivers an X-ray dose of approximately 5 × 10^8 Gy. The sample is stable for the first two exposures but then experiences a rapid collapse, followed by a slow but continuous shrinkage. The collapse is apparent from the elongated speckles, which indicate a sample for which the diameter is changing during the exposure while its relative structure is maintained. Overall, the sample loses about 25% of its volume prior to the final exposure used for reconstruction (Shapiro *et al.*, 2005; Thibault *et al.*, 2006). Over the course of this exposure series the total scattered signal does not change, indicating that the total mass of the sample remains intact. The rapid change in the diffraction pattern during the early exposures means that successful imaging experiments require pre-irradiation of the sample. The slow but continuous shrinkage of dry samples with further dose means that the resolution of the three-dimensional images will necessarily be reduced. Indeed, Nishino *et al.* (2009) observed reduced resolution in their three-dimensional reconstruction of a dry chromosome because of morphological changes which occurred during data collection. Hence, full three-dimensional imaging of cells by diffractive methods requires cryogenic protection against radiation damage. Predictions based on a calculation of the cross section for coherent scattering by a smooth dielectric indicate that 10 nm resolution imaging of frozen hydrated organic matter should be possible using soft X-rays at currently available synchrotron sources (Howells *et al.*, 2009). This limit is arrived at through a comparison of the radiation dose required for imaging and the dose at which radiation damage has been empirically observed at different length scales. However, it seems plausible that the presence of many identical particles within a cell could be exploited to provide super-resolution information.

9.3.3.2. Low-dose three-dimensional imaging; low damage potential of stereoscopic viewing

Diffraction imaging in three dimensions proceeds as it does in standard X-ray tomography. That is, two-dimensional data are recorded from many angular orientations of the sample and then assembled into a three-dimensional data set. In this case, however, the data are recorded in reciprocal space and the individual data sets only need to be registered with respect to the angular coordinate due to the Fourier shift theorem. In the absence of any additional information, the angular sampling of reciprocal space is determined by the Crowther resolution,

$$k_C = \frac{1}{\Delta\theta D},$$

where D is the object diameter and $\Delta\theta$ is the angular separation of the two-dimensional data sets. This is the spatial frequency at which the unmeasured Fourier components, those in between the measured Ewald sphere segments, can be properly interpolated from the measured data. Diffraction microscopy, however, requires the addition of information in the form of real-space constraints. This additional information allows for the calculation of not only the missing reciprocal-space phases but also a limited number of missing magnitudes. Chapman *et al.* (2006) showed that $\Delta\theta$ could in fact be up to four times larger than required by the Crowther relation, with k_C matching the numerical aperture of the imaging system. Thus, three-dimensional reconstructions could take place with nearly isotropic diffraction-limited resolution with only about 150 angular orientations of the sample.

Stereoscopic viewing can provide a significant degree of three-dimensional perception of an extended object while only increasing the total radiation exposure by a factor of two. In principle, according to the dose-fractionation theorem of Hegerl and Hoppe, full three-dimensional visualization of a given resolution element should not require a dose any higher than two-dimensional visualization of the same element with the same statistical accuracy (Hegerl & Hoppe, 1976). This theorem provides hope that high-resolution imaging in three-dimensions, perhaps even of dry specimens, is possible, but in practice this is very difficult to achieve and low-dose imaging techniques are only now being explored by the CXDM community. Stereoscopic viewing should be considered the preliminary low-dose technique of choice. One particular advantage is the rapid reconstruction (compared with full three-dimensional reconstructions) which makes possible *in situ* sample inspection. Fig. 9.3.3.3 shows a stereo image of a chemically dried budding yeast cell. When viewed stereoscopically, with the viewer's focus in front of the image, the three-dimensional arrangement of a group of vesicles in the mother cell can be visualized.

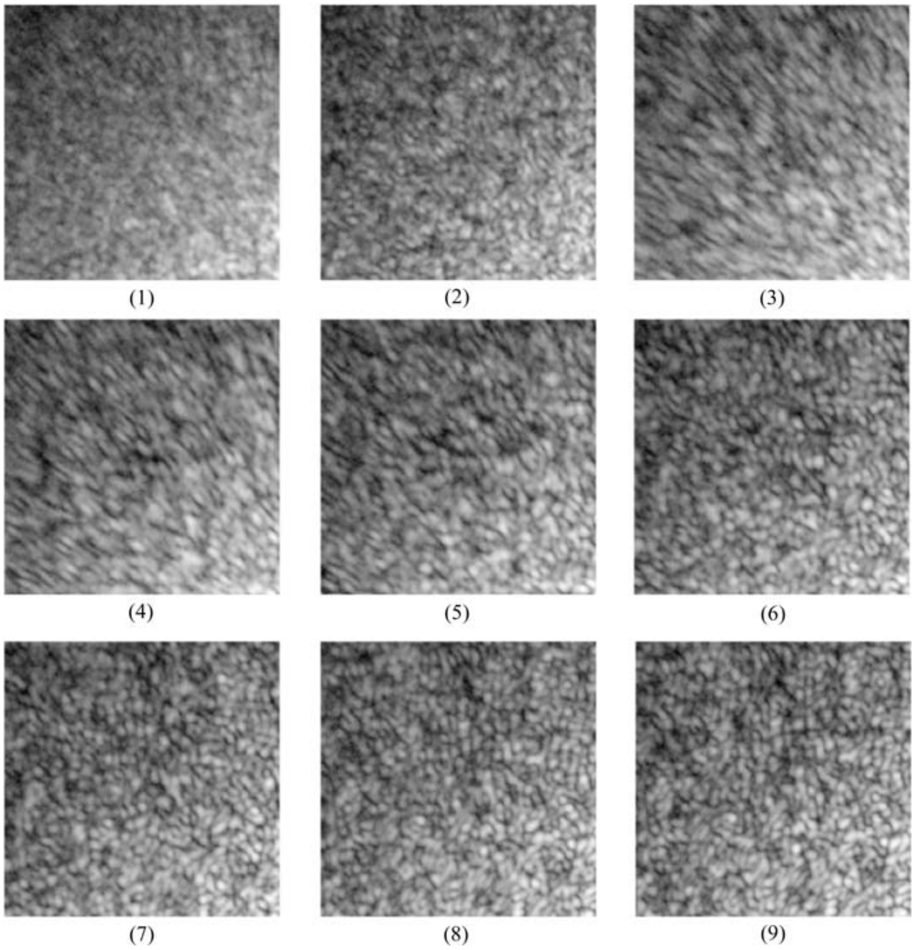


Figure 9.3.3.2
Exposure to ionizing radiation causes shrinkage of organic matter. Each image in this series is a section of a measured diffraction pattern from a freeze-dried yeast cell. The images were taken sequentially and each represents an additional X-ray dose of 5×10^8 Gy. After a cumulative dose of 1×10^9 Gy, the cell undergoes a rapid collapse [apparent from the elongated speckles in images (3)–(5)] followed by continued shrinkage at a reduced rate. The X-rays used had an energy of 750 eV and a dose of 5×10^8 Gy was adequate for reconstruction at 30 nm resolution. (Reproduced from Shapiro, 2004).

9.3.4. Conclusions

CXDM promises to be a highly efficient imaging methodology which can deliver high-resolution and high-contrast images of large non-crystalline biological structures. Radiation-induced shrinkage of dry cells will probably prohibit three-dimensional imaging of such cells at high resolution. However, significant three-dimensionality can be achieved through stereoscopic viewing of a cell, which only doubles the necessary X-ray dose. Even so, in the absence of low-dose diffraction techniques the sample must undergo considerable morphological change prior

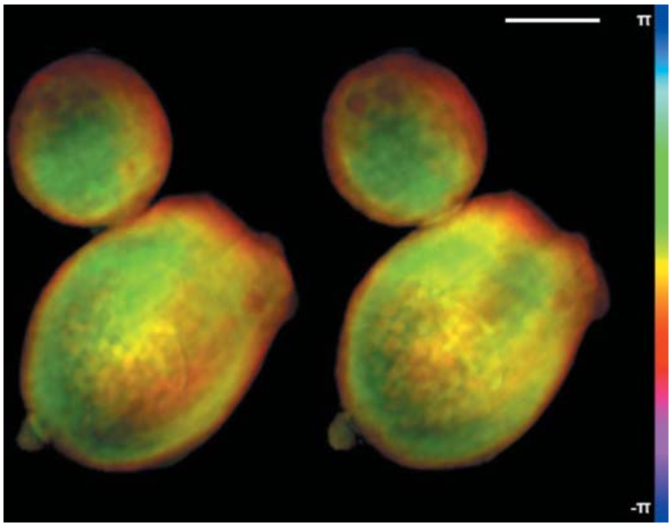


Figure 9.3.3.3
Stereo image of a budding yeast cell. This budding yeast cell was chemically fixed with gluteraldehyde and dehydrated in acetone. The images have an angular separation of 10° and a pixel size of 11 nm. The three-dimensional arrangement of a group of small vesicles in the mother cell can be visualized when viewed stereoscopically (with the viewer's focus in front of the image). The scale bar is 500 nm.

to imaging, although this change does not seem to alter the relative arrangement of organelles. The development of low-dose techniques will allow for the direct observation of these radiation-induced changes. In the long run, it is cryogenic protection that provides the most valuable structural information, since the cells are maintained in a near living state. Ultra-high-resolution three-dimensional imaging will still require the development of low-dose techniques, as cryoprotected samples have also been observed to suffer from mass loss with the accumulation of very high X-ray doses ($>10^{10}$ Gy). These imaging techniques are currently under development at the ALS in collaboration with Stony Brook University and elsewhere. Alternatively, X-ray free-electron lasers (FELs) promise the highest resolution imaging of living cells that is possible by any means; indeed, it has been proposed that sub-nanometre resolution is possible (Bergh *et al.*, 2008). The ultra-short and ultra-bright pulses of an X-ray FEL will encode the structural information from a living cell before it is destroyed by the pulse.

The author would like to thank Johanna Nelson and Xiaojing Huang of Stony Brook University for preparing the samples and providing the diffraction data for reconstruction. This work was supported under the Seaborg Fellowship at Lawrence Berkeley National Laboratory and by grants from the National Institutes of Health, the National Science Foundation and the US Department of Energy to Stony Brook University. The Advanced Light Source at Lawrence Berkeley National Laboratory is supported by the Director, Office of Science, Office of Basic Energy Sciences, Materials Sciences Division, of the US Department of Energy.

References

Barty, A., Marchesini, S., Chapman, H. N., Cui, C., Howells, M. R., Shapiro, D. A., Minor, A. M., Spence, J. C., Weierstall, U., Ilavsky, J., Noy, A., Hau-Riege, S. P., Artyukhin, A. B., Baumann, T., Willey, T., Stolken, J., van Buuren, T. & Kinney, J. H. (2008). *Three-dimensional coherent X-ray diffraction imaging of a ceramic nanofoam: determination of structural deformation mechanisms. Phys. Rev. Lett.* **101**, 055501.

Bates, R. H. T. (1982). *Fourier phase problems are uniquely soluble in more than one dimension. I. Underlying theory. Optik (Stuttgart)*, **61**, 247–262.

Bates, R. H. T. & Fright, W. R. (1983). *Composite two-dimensional phase-restoration procedure. J. Opt. Soc. Am.* **73**, 358–365.

Bergh, M., Huldt, G., Timneanu, N., Maia, F. R. & Hajdu, J. (2008). *Feasibility of imaging living cells at subnanometer resolutions by ultrafast X-ray diffraction. Q. Rev. Biophys.* **41**, 181–204.

Chao, W., Kim, J., Rekawa, S., Fischer, P. & Anderson, E. H. (2009). *Demonstration of 12 nm resolution Fresnel zone plate lens based soft X-ray microscopy. Opt. Express*, **17**, 17669–17677.

Chapman, H. N., Barty, A., Marchesini, S., Noy, A., Hau-Riege, S. P., Cui, C., Howells, M. R., Rosen, R., He, H., Spence, J. C., Weierstall, U., Beetz, T., Jacobsen, C. & Shapiro, D. (2006). *High-resolution ab initio three-dimensional X-ray diffraction microscopy. J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **23**, 1179–1200.

Fienup, J. R. (1978). *Reconstruction of an object from the modulus of its Fourier transform. Opt. Lett.* **3**, 27–29.

Fienup, J. R. (1982). *Phase retrieval algorithms: a comparison. Appl. Opt.* **21**, 2758–2769.

Fienup, J. R. (1987). *Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **4**, 118–123.

Fienup, J. R., Crimmins, T. R. & Holsztynski, W. (1982). *Reconstruction of the support of an object from the support of its autocorrelation. J. Opt. Soc. Am.* **72**, 610–624.

Gerchberg, R. W. & Saxton, W. O. (1972). *A practical algorithm for the determination of phase from image and diffraction plane pictures. Optik (Stuttgart)*, **35**, 237–246.

Hegerl, R. & Hoppe, W. (1976). *Influence of electron noise on three-dimensional image reconstruction. Z. Naturforsch. Teil A*, **31**, 1717–1721.

Howells, M. R., Beetz, T., Chapman, H. N., Cui, C., Holton, J. M., Jacobsen, C. J., Kirz, J., Lima, E., Marchesini, S., Miao, H., Sayre, D., Shapiro, D. A., Spence, J. C. H. & Starodub, D. (2009). *An assessment of the resolution limitation due to radiation damage in X-ray diffraction microscopy. J. Electron Spectrosc. Relat. Phenom.* **170**, 4–12.

Huang, X., Nelson, J., Kirz, J., Lima, E., Marchesini, S., Miao, H., Nieman, A. M., Shapiro, D., Steinbrener, J., Stewart, A., Turner, J. J. & Jacobsen, C. (2009). *Soft X-ray diffraction microscopy of a frozen hydrated yeast cell. Phys. Rev. Lett.* **103**, 198101.

Jearanaikoon, S. & Abraham-Peskir, J. V. (2005). *An X-ray microscopy perspective on the effect of glutaraldehyde fixation on cells. J. Microsc.* **218**, 185–192.

Larabell, C. A. & Le Gros, M. A. (2004). *X-ray tomography generates 3-d reconstructions of the yeast, Saccharomyces cerevisiae, at 60 nm resolution. Mol. Biol. Cell*, **15**, 957–962.

Le Gros, M. A., McDermott, G., Uchida, M., Knoechel, C. G. & Larabell, C. A. (2009). *High-aperture cryogenic light microscopy. J. Microsc.* **235**, 1–8.

Lima, E., Wiegart, L., Pernot, P., Howells, M. R., Timmins, J., Zontone, F. & Madsen, A. (2009). *Cryogenic X-ray diffraction microscopy for biological samples. Phys. Rev. Lett.* **103**, 198102.

Marchesini, S. (2007). *A unified evaluation of iterative projection algorithms for phase retrieval. Rev. Sci. Instrum.* **78**, 011301.

Marchesini, S., He, H., Chapman, H. N., Hau-Riege, S. P., Noy, A., Howells, M. R., Weierstall, U. & Spence, J. C. H. (2003). *X-ray image reconstruction from a diffraction pattern alone. Phys. Rev. B*, **68**, 140101.

Miao, J., Charalambous, P., Kirz, J. & Sayre, D. (1999). *Extending the methodology of X-ray crystallography to allow imaging of micrometre-sized non-crystalline specimens. Nature (London)*, **400**, 342–344.

Miao, J., Ishikawa, T., Johnson, B., Anderson, E. H., Lai, B. & Hodgson, K. O. (2002). *High resolution 3D X-ray diffraction microscopy. Phys. Rev. Lett.* **89**, 088303.

Miao, J., Sayre, D. & Chapman, H. N. (1998). *Phase retrieval from the magnitude of the Fourier transforms of nonperiodic objects. J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **15**, 1662–1669.

Nelson, J., Huang, X., Steinbrener, J., Shapiro, D. A., Kirz, J., Marchesini, S., Neiman, A. M., Turner, J. & Jacobsen, C. (2010). *High resolution X-ray diffraction microscopy of speciffically labelled yeast cells. Proc. Natl Acad. Sci. USA*, **107**, 7235–7239.

Nishino, Y., Takahashi, Y., Imamoto, N., Ishikawa, T. & Maeshima, K. (2009). *Three-dimensional visualization of a human chromosome using coherent X-ray diffraction. Phys. Rev. Lett.* **102**, 018101.

Pfeifer, M. A., Williams, G. J., Vartanyants, I. A., Harder, R. & Robinson, I. K. (2006). *Three-dimensional mapping of a deformation field inside a nanocrystal. Nature (London)*, **442**, 63–66.

Sayre, D. (1980). *Imaging Processes and Coherence in Physics*, edited by M. Schlenker, M. Fink, J. P. Goedgebuer, C. Malgrange, J. Ch. Viénot & R. H. Wade, p. 229. Heidelberg: Springer-Verlag.

Sayre, D. (1991). *Direct Methods of Solving Crystal Structures*, edited by H. Schenk, pp. 353–355. New York: Plenum Press.

Sayre, D. (2008). *Report on a project on three-dimensional imaging of the biological cell by single-particle X-ray diffraction. Acta Cryst.* A**64**, 33–35.

Sayre, D. & Chapman, H. N. (1995). *X-ray microscopy. Acta Cryst.* A**51**, 237–252.

Schroer, C. G., Boye, P., Feldkamp, J. M., Patommel, J., Schropp, A., Schwab, A., Stephan, S., Burghammer, M., Schöder, S. & Riekel, C. (2008). *Coherent X-ray diffraction imaging with nanofocused illumination. Phys. Rev. Lett.* **101**, 090801.

Shapiro, D., Thibault, P., Beetz, T., Elser, V., Howells, M., Jacobsen, C., Kirz, J., Lima, E., Miao, H., Neiman, A. M. & Sayre, D. (2005). *Biological imaging by soft X-ray diffraction microscopy. Proc. Natl Acad. Sci. USA*, **102**, 15343–15346.

Shapiro, D. A. (2004). *Biological Imaging by Soft X-ray Diffraction Microscopy*. PhD thesis, Stony Brook University.

Thibault, P., Elser, V., Jacobsen, C., Shapiro, D. & Sayre, D. (2006). *Reconstruction of a yeast cell from X-ray diffraction data. Acta Cryst.* A**62**, 248–261.

Wang, Y., Jacobsen, C., Maser, J. & Osanna, A. (2000). *Soft X-ray microscopy with a cryo scanning transmission X-ray microscope: II. Tomography. J. Microsc.* **197**, 80–93.

Weiss, D., Schneider, G., Niemann, B., Guttman, P., Rudolph, D. & Schmahl, G. (2000). *Computed tomography of cryogenic biological specimens based on X-ray microscopic images. Ultramicroscopy*, **84**, 185–197.

International Tables for Crystallography (2012). Vol. F, Chapter 10.1, pp. 241–248.

PART 10. CRYOCRYSTALLOGRAPHY

Chapter 10.1. Introduction to cryocrystallography

H. Hope and S. Parkin

Since about 1985, low-temperature methods in biocrystallography have evolved from stumbling experimentation to mainstay production techniques. This would not have happened without good reason. Most of the early development was done with nitrogen cooling. More recently, temperatures as low as 5 K have become accessible by means of liquid-helium-based techniques. A brief discussion of the methods, equipment and advantages of data collection at cryogenic temperatures is given.

10.1.1. Cooling of biocrystals

10.1.1.1. *Physical chemistry of biocrystals*

Crystals are normally brought from room temperature to the working, low temperature by relatively rapid cooling, either in a cold gas stream, or by immersion in a cryogen such as liquid nitrogen or liquid propane. One goal of the procedure is to avoid crystallization of any water present in the system, whether internal or external to the crystal. Ice formation depends on the formation of nuclei. Nuclei are formed either by homogenous nucleation, *i.e.* in bulk liquid, or by heterogeneous nucleation, *i.e.* at the surface of a phase other than the liquid. Although data pertaining to biocrystals are scarce, indications are that internal nucleation, whether homogenous or heterogeneous, is not common. Proteins that induce nucleation at mild supercooling are known, so presumably there exist regions in these proteins that help to prearrange water molecules so that they readily form ice nuclei. There are also proteins that hinder nucleation. At present there is no basis for predicting the outcome of cooling for any given protein crystal. Only in a statistical sense can one be reasonably confident that a given macromolecule will not promote the freezing of water.

Vali and coworkers (Götz *et al.*, 1991; Vali, 1995) have provided a quantitative treatment of ice nucleation that can serve as a guideline. They observe that the absolute rate of formation of nuclei increases with the volume of water and with decreasing temperature. The probability p that a volume V of water will begin freezing during a time span t is given by

$$p = J(T)Vt,$$

where $J(T)$ is the nucleation rate at temperature T. Based on empirical data, $J(T)$ is given by

$$J(T) = 6.8 \times 10^{-50} \exp[3.9(273 - T)],$$

where J is in $\mathrm{m^{-3}\,s^{-1}}$ and T is in K. Note that $J(T)$ increases by a factor of 50 per K. As a practical limit, bulk water cannot be cooled below 233 K without freezing. However, given a sufficiently small volume and high cooling rate, it is possible to supercool water to form a glassy state that is at least kinetically stable. Stability requires a temperature below 140 K; at higher temperatures crystallization eventually takes place. For the cooling rates typically attained with small crystals (up to a few hundred K $\mathrm{s^{-1}}$) it seems impossible to avoid crystallization of water in the mother liquor adhering to a crystal, unless it is modified in some way. Once ice forms at the crystal surface, freezing may propagate through the entire crystal, effectively destroying it. Even if the crystal remains intact, diffraction from polycrystalline ice will render parts of any data set from that crystal useless. Because the probability of a nucleation event increases with time, it seems prudent to use a rapid cooling process. However, we note that the expression for $J(T)$ is formulated for pure water and cannot be valid for all conditions; it is well established that a majority of biocrystals can be cooled below 140 K.

A consequence of the foregoing is that for prevention of ice growth one should first focus attention on the region immediately outside the crystal, rather than on its interior. Two approaches have been shown to have merit: (*a*) modification of the solvent layer, and (*b*) removal of the solvent layer.

The goal of solvent modification is to prevent ice formation in that layer. Commonly used modifiers (referred to as antifreezes or cryoprotectants) are water-soluble organic compounds of low molecular weight with good hydrogen-bonding properties; examples include glycerol, ethylene glycols and MPD (2-methylpentane-2,4-diol). These compounds are added to reach a concentration sufficient to suppress nucleation and thereby prevent ice formation. Typical concentrations are in the 15–30% range, depending on the compound and the original composition of the mother liquor. The required concentration must be determined by experiment. Some suitable starting points are given by Garman & Mitchell (1996). The modified solution is tested by cooling a small drop to the working temperature. If the drop remains clear, there is no ice formation.

It is important to keep in mind that any change in the properties of the medium surrounding the crystal will have consequences for its crystallographic stability. In order to protect the crystal, two fields should be considered: thermodynamics and kinetics.

For a crystal in equilibrium with its mother liquor, the chemical potential of each species will be the same inside the crystal and in the mother liquor. If the solution surrounding the crystal is altered by the addition of an antifreeze, the chemical potential μ of water (and other species) will change and the crystal will no longer be in equilibrium with its surrounding solution. The typical result is that $\mu(\mathrm{H_2O}, \text{solution})$ decreases, so $\mu(\mathrm{H_2O}, \text{crystal}) > \mu(\mathrm{H_2O}, \text{solution})$ and there will be a thermodynamic drive to remove water from the crystal. The activation energy for water diffusion is low, so if the process is allowed to proceed, the end result is loss of water with likely deterioration in crystal quality (but see below). Considerations of this kind led Schreuder *et al.* (1988) to develop procedures for solvent modification that would prevent destruction of the crystal. Although some success was reported, sufficient problems were encountered that the approach cannot be considered to be a general solution.

Copyright © 2012 International Union of Crystallography

It is important to note that loss of water does not always lead to loss of crystal integrity. For example, Esnouf *et al.* (1998) and Fu *et al.* (1999) have shown that controlled dehydration can result in substantially improved resolution. In addition, antifreeze concentrations much higher than those needed to suppress ice formation (Mitchell & Garman, 1994) can preserve low mosaic spread. Work by Kriminski *et al.* (2002) suggests these phenomena may be connected.

In earlier work, Travers & Douzou (1970) emphasized the importance of keeping the dielectric constant unchanged when modifying the mother liquor. Petsko (1975) made observations that support the significance of this approach and, based on systematic studies, also showed that keeping $\mu(\mathrm{H}^+)$ constant is of great importance. Hui Bon Hoa & Douzou (1973) and Douzou *et al.* (1975) have presented tables of solvent compositions that facilitate the preparation of successful cryoprotective solutions. It should be noted that a significant aim in Petsko's work was to keep the solvent liquid, so as to permit manipulation of enzyme substrates. Studies of enzyme kinetics are much more demanding than the rapid cooling to about 100 K that is of primary interest here.

In most cases it is only necessary to consider kinetic effects, *i.e.*, how long it takes before the crystal itself begins to change. When a crystal in a drop of its original mother liquor is dipped into a drop of modified mother liquor, diffusion begins immediately. The speed of propagation in the liquid phase can be estimated from a standard equation for the mean-square travel distance of a diffusing species,

$$\overline{x^2} = 2Dt,$$

where D is the diffusion coefficient and t is the the time. Typical room-temperature values for D for antifreeze molecules in water are around 10^{-9} m^2 s^{-1}. Thus, a root-mean-square travel distance of 0.1 mm requires about 5 s. For a solvent layer about 0.1–0.2 mm thick, a contact time of 5–20 s should provide a sufficient level of modification to prevent freezing, while the risk of crystal damage is small. It is often important to stop any ongoing process as soon as protection from freezing has been attained. This can conveniently be achieved by immersion in liquid N_2.

10.1.1.2. Internal ice or phase transition

If there are good indications that ice formation does start internally, or that a destructive phase transition takes place, an attempt can be made to modify the internal water structure. An important consideration of Petsko (1975) was never to allow large deviations from equilibrium. This can be accomplished by a slow, gradual change in $\mu(\mathrm{H_2O}$, solution), allowing enough time for the crystal to re-establish equilibrium. A number of successful experiments were reported.

10.1.1.3. Removal of the solvent layer

Because of their tendency toward rapid loss of internal solvent, biocrystals rarely survive prolonged exposure to the atmosphere. A solution to this problem was described by Hope (1988), where the solvent is removed while the crystal is submerged in a hydrocarbon oil. After the liquid has been removed, a small drop of oil is allowed to encapsulate the crystal, allowing it to tolerate brief exposure to air. Even under such mild conditions, some crystals still lose water and suffer damage. A remedy for this is to keep the oil saturated with water. One disadvantage of the oil technique is the tendency of loop mounts to carry along too much oil (Teng, 1990), which can cause excessive background scattering. An advantage is that absorption can become nearly isotropic. The most commonly used oil is the polyisobutene Infineum V8512, formerly known as Infineum Parabar 10312, Exxon Paratone-8277 or Paratone-N. Contrary to popular myth, there is nothing magical or mysterious about this particular oil. Important properties are that it is inert, has a useful viscosity, forms a glass on cooling and has a coefficient of thermal expansion which appears to match that of many biocrystals.

10.1.1.4. Cooling rates

The time dependence of nucleation probability suggests that faster is safer. Although few systematic data are available, it is commonly assumed that crystal cooling should be as rapid as possible. Studies related to cryopreservation of biological samples for electron microscopy provide a number of measurements of cooling rates in various coolants, but it is difficult to extract information directly relevant to cryocrystallography. From a practical point of view, the coolants to be considered are liquid N_2 and liquid propane (and, to a lesser extent, liquid ethane). Thermal conductivities for small-molecule compounds in liquid form tend to be of similar magnitude – around 1.5×10^{-5} W m^{-1} K^{-1}. N_2 boils at 77 K; propane remains liquid between 83 and 228 K. It is often thought that the gas layer that can form around an object dipped in liquid N_2 as a result of the Leidenfrost effect (Leidenfrost, 1756) makes liquid N_2 less effective as a coolant than liquid propane, which is much less likely to form bubbles. However, from model calculations, Bald (1984) suggested that this Leidenfrost insulation problem in liquid N_2 would not be significant in the cooling of small objects of low thermal conductivity, because there is not enough heat transport to the surface to maintain the gas layer. He also concluded that liquid N_2 could potentially yield the highest cooling rate among commonly used coolants, but in a review of plunge-cooling methods, Ryan (1992) gives preference to liquid ethane. Walker *et al.* (1998) measured the cooling rates in N_2 gas (100 K), liquid N_2 (77 K) and liquid propane (100 K) of a bare thermocouple and of a thermocouple coated with RTV silicone cement. The thermocouples were made from 0.125-mm wire and the coating was about 0.20–0.25 mm thick. With the gas stream, cooling of the centres of the samples from 295 to 140 K took 0.8 and 2 s, respectively; with liquid N_2 the times were 0.15 and 0.6 s, and with liquid propane they were 0.15–0.18 and 1.2 s (time reproducibility is to within ±10%). Given the simplicity of liquid-N_2 immersion, there seems little reason to choose the more complicated and more hazardous liquid-propane technique. As the field of low-temperature biocrystallography has matured, liquid-propane methods have all but died out, and liquid-N_2 immersion is now by far the most commonly employed method.

10.1.2. Beneficial effects of low temperature

10.1.2.1. Suppression of radiation damage

Biocrystals near room temperature are sensitive to X-rays and generally suffer radiation damage during data measurement. Often this damage is so rapid and severe that a number of different crystals are needed for a full data set. On occasion, damage is so rapid that data collection is impossible. Crystal decay is typically accompanied by changes in reflection profiles and cell dimensions, which alter the positions of diffraction maxima, exacerbating the problem of changing diffraction intensities. The use of more than one crystal invariably introduces inaccuracies. Intensities from a crystal near the end of its usable life will have decay errors. Individual samples of biocrystals frequently have measurable differences in structure; merging of data will result in an average of the structures encountered, with concomitant loss of definition. Crystals cooled to near liquid-N_2 temperature typically show a greatly reduced rate of radiation damage, often to the extent that it is no longer an issue of major concern. The protection from radiation damage was noted early on (Haas & Rossman, 1970) and numerous cases were observed by Petsko (1975). A noteworthy example is the successful prevention of radiation damage to crystals of ribosome particles (Hope *et al.*, 1989).

Radiation damage appears to be most commonly initiated by photoelectrons and propagated by their inelastic scattering from nearby atoms, creating positively charged species and a cascade of secondary electrons (Garman & Nave, 2009). These can further interact with either protein or with water to form reactive, charged species and free radicals. At sufficiently low temperature, two effects can influence the rate of damage: movement of the reactive species is impeded and the activation energy for reaction is not available. A revealing observation has been described by Hope (1990), where a crystal that had been exposed to synchrotron radiation for many hours at 85 K showed no overt signs of radiation damage, but as the crystal was being warmed toward room temperature, it suddenly turned black and curled up like a drying leaf. More commonly, crystals turn yellow under X-ray irradiation, and bubbles and cracks appear on warming. The rate of free-radical formation would be little affected by temperature, so that when sufficient mobility and activation energy become available, the stored radicals will react.

There is a general consensus that radiation damage, even in high-flux synchrotron beams, can be slowed by cooling to liquid-helium temperatures. The extent of radiation damage is expected to depend on the nature of the macromolecule, and literature examples clearly illustrate this. Meents *et al.* (2007) have reported an extensive study of insulin and holoferritin at 15 and 90 K. They observed statistically significant but relatively small reductions in radiation damage; for holoferritin, 23% less damage at the lower temperature, and for insulin about 6%. In a study of *Streptomyces rubiginosus* D-xylose isomerase, Chinte *et al.* (2007) found a lifetime extension of 25% at 8 K compared to data collection at 100 K. Corbett *et al.* (2007) compared results from data collected at ~40 and at 110 K in a study of the metalloprotein putidaredoxin. They report that radiation-induced photoreduction at 110 K resulted in misleading structure interpretation. At 40 K the photoreduction did not occur. The damage mechanism in this case is related to a change in oxidation state of a central metal atom. The authors made a strong argument for measuring data for metalloproteins at the lowest possible temperature. Chinte *et al.* (2007) pointed out that the additional cost of liquid-helium-temperature measurements compared to 90–100 K measurements is small, and that the advantages can significantly outweigh the increased cost.

In recent years, the ability to calibrate X-ray beam intensity, to calculate reasonable approximations to accumulated radiation dose and to assess crystal degradation by means of various spectroscopies (*e.g.* UV–visible, IR, Raman, XAFS, XANES) in concert with diffraction has brought dramatic advances in understanding radiation damage by X-rays. A concise account has been given by Garman & Nave (2009). Useful suggestions for avoidance of radiation damage, including cases where multiple crystals are required, have been given by Holton (2009). Briefly, knowledge of beamline photon flux density (photons $\mu m^{-2}\, s^{-1}$), the dose ratio (a function of crystal composition and X-ray wavelength), the crystal shape and the cross section intensity profile of a given X-ray beam allows an estimate of the maximum tolerable exposure time for a particular crystal.

10.1.2.2. Mechanical stability of the crystal mount

The mechanical stability of samples is also of concern. Crystals mounted in capillaries and kept wet are prone to movement, giving rise to difficulties with intensity measurements. A crystal at cryotemperature is rigidly attached to its mount; slippage is impossible.

10.1.2.3. Effect on resolution

The effects on radiation damage and mechanical stability are clear-cut, and provide the main reasons for using cryotechniques. Resolution can also be affected, but the connection between temperature and resolution is neither simple nor obvious.

In any crystal, the Boltzmann distribution law is an important factor in determining the accuracy of the replication of structure from one unit cell to another. For many small-molecule crystals, just one arrangement corresponds to a distinct energy minimum. The result is a well ordered structure. With macromolecules, the typical situation is one where a number of arrangements correspond to similar energies. Accordingly, a number of atomic arrangements will be expressed in the crystal. Although the relative values of local minima depend on the temperature, one cannot count on a significant change in ordering by cooling the crystal. Instead, some distribution will be frozen in.

If poor resolution is the result of rapid radiation damage, data collection at cryotemperature can lead to much improved resolution. However, if poor resolution is caused mainly by inexact replication from one unit cell to another, lowering the temperature may have little effect on resolution. If the mosaic spread in the crystal increases upon cooling, resolution may even deteriorate.

In a model proposed by Hope (1988), a relationship between resolution r and temperature T is given by

$$r_2 = r_1[(B_0 + bT_2)/(B_0 + bT_1)]^{1/2}.$$

Here r_1 is the resolution at T_1, r_2 is the resolution at T_2, B_0 is the value of B at $T = 0$ and b is a proportionality constant. There are two underlying assumptions: (1) the overall atomic distribution does not change significantly with temperature and (2) for any given T, the temperature factor [*i.e.* $\exp(-B\sin^2\theta/\lambda^2)$] at the resolution limit has the same value; thus the effects of scattering factors and Lorentz–polarization factors are ignored. We see that if B_0 is the predominant term, lowering T will not have much effect, whereas for small B_0 (a relatively well ordered structure) the effect of T on r can be large. For example, if the room-temperature resolution is 1.5 Å, the resolution at 100 K can be around 1 Å, but if the room-temperature resolution is around 3 or 4 Å, little change can be expected. A qualitative assessment of these effects was clearly stated by Petsko (1975).

10.1.2.4. Annealing of biocrystals

Prior to about 1996, it was thought that thawing of a flash-cooled crystal would inevitably lead to its demise. In spite of anecdotal evidence that some biocrystals could survive warming and re-cooling, this notion persisted until work by Harp *et al.* (1998) and by Yeh & Hol (1998) showed that some biocrystals could be annealed under certain well defined conditions. The method of Harp *et al.* (1998) involved transfer of a flash-cooled crystal from the cold stream to a cryoprotective solution at room temperature for 3 min followed by a second flash cooling. The Yeh & Hol (1998) technique, on the other hand, is performed *in situ* by simply blocking the cold stream for 1–2 s (*i.e.* until melting is observed), after which time the blockage is removed to re-cool the sample. Both these annealing protocols were shown to be capable of dramatic improvements in diffraction quality, both in terms of reduced mosaicity and improved resolution. A plausible mechanism involving the release of cooling-induced lattice stress by defect migration and solvent transport was suggested by Kriminski *et al.* (2002). Other work (Parkin & Hope, 2003; Juers & Matthews, 2004; Weik *et al.*, 2005) supports the notion of solvent transport, possibly as a result of solvent crystallization (Weik *et al.*, 2001) or other phase transition (Parkin & Hope, 2003) in the aqueous regions within biocrystals.

10.1.2.5. Additional benefits from sub-77 K cooling with helium

As is the case with nitrogen cooling, it is unlikely that thermodynamic equilibrium will be reached by cooling to liquid-helium temperatures. The change that will certainly take place on cooling to liquid-helium temperature is that true thermal motion will be greatly reduced. One result of this is that individual atom peaks will become much sharper. For example, electron-density maxima for well ordered atoms will increase by a factor of about three on cooling from 90 to 10 K. Potentially, this can allow a more detailed interpretation of a structure with a resolution limit better than about 1.5 Å, and also for the better ordered regions of a structure with poorer overall resolution. In general, however, it is not realistic to expect a significant resolution improvement in low-resolution structures based on the effects of temperature alone. Improvements related to diminished radiation damage, on the other hand, can be significant. Two studies illustrate the effects discussed here.

The effects of helium cooling on a high-resolution structure are well illustrated in a study by Petrova *et al.* (2006). They studied a complex of human aldose reductase at 15, 60 and 100 K. The complex has yielded data to 0.66 Å resolution, and thus represents a generally highly ordered structure. The emphasis of the study was on the behaviour of the atomic displacement parameters (ADPs). It was found that the major ADP component for well ordered atoms is temperature driven, as it would be in normal small-molecule structures. A large proportion of the atoms at 15 K have B values of 2 Å^2 or less (about 0.025 Å^2 or less in terms of U values). At 100 K, the corresponding cutoff is about 5 Å^2. Cooling to 15 K allows large portions of the structure to be determined with a precision that would be considered excellent for small molecules. However, the average isotropic B value for the 'best' Cα atoms at 15 K is still 3.9 Å^2 (a U value of about 0.05 Å^2). This indicates that the positional parameters for many of these atoms in reality are composites of closely spaced positions. The best average protein ADPs at 15 K are about the magnitude of small-molecule ADPs at room temperature. This sets unfortunate limits to the attainable accuracy of structural and electron-density parameters.

Hexagonal hen egg-white lysozyme has a relatively well ordered structure, but there are significant regions with multiple conformations. Brinkmann *et al.* (2006) measured diffraction data at 10 K to a resolution limit of 1.46 Å. The results indicate that major areas of disorder are present, illustrating that structural disorder persists at the lowest temperatures.

Although helium is more expensive than nitrogen as a coolant, the added cost for a helium-temperature data set is usually trivial. Equipment design and operating methods have developed to a stage where there is no significant operational difference between nitrogen and helium cooling when manual crystal handling is used.

10.1.3. Principles of cooling equipment

There are many ways to construct a low-temperature apparatus based on the cold-stream principle that functions well, but they are all made according to a small number of basic principles.

All gas-stream crystal-cooling devices must have three essential components: (*a*) a cold gas supply, (*b*) a system of cold gas delivery to the crystal, and (*c*) a system for frost prevention at the crystal site.

10.1.3.1. Liquid-nitrogen-based cold gas supply

Historically, two methods were commonly used: generation of gas by boiling liquid N_2 with an electrical heater, and cooling of a gas stream in a liquid-N_2 heat exchanger. The currently common methods are boiling, and cooling of the gas by means of a refrigerator.

Because precise voltage and current control are easily realized, the boiler method has the advantage of providing very accurate control of the flow rate with minimal effort. Precise control of the flow rate is typically not attained when the rate is controlled with standard gas-flow regulators, because they control volume, not mass.

In addition to control of the flow rate, precise control of the temperature requires exceptional insulation for the cold stream. The longer the stream path, the higher the requirements for insulation. As a rule, temperature rise during transfer should not exceed 15 K at a flow rate of 0.2 mol N_2 min^{-1}; preferably, it should be significantly lower. Higher cooling loss leads to excessive coolant consumption and to instability caused by changes in ambient temperature. High flow rates may also tend to cause undesirable cooling of diffractometer parts.

Appropriate insulation can be readily attained either with silvered-glass Dewar tubing or with stainless-steel vacuum tubing. Glass has the advantage of being available from local glassblowing shops; it generally provides excellent insulation. The main disadvantages are fragility and a rigid form that makes accurate positioning of the cold stream difficult. Stainless steel can provide superb insulation, given an experienced manufacturer. A major advantage is the availability of flexible transfer lines that greatly simplify the positioning of the cold stream relative to the diffractometer.

10.1.3.2. Liquid-helium-based cold gas supply

Open-stream cooling devices that can reach temperatures around 5 K are now commercially available. In principle, the design is simpler than that for liquid-nitrogen-based devices. A basic cooling apparatus consists of a liquid-helium transfer line from a pressurized delivery tank, an evaporation chamber and a cold gas delivery tube. The transfer line is an insulated capillary tube. The delivery tube is an insulated stretch of vacuum tubing with an electrically heated nozzle at the exit. The helium flow is controlled with a needle valve. Because of thermal loss, the flow rate also largely determines the temperature in the range below 20 K. Above about 20 K, an in-stream heating element is used for additional temperature control. This is necessary, because if the flow rate is too low, the cooling stream becomes unstable and will not reliably cover the sample.

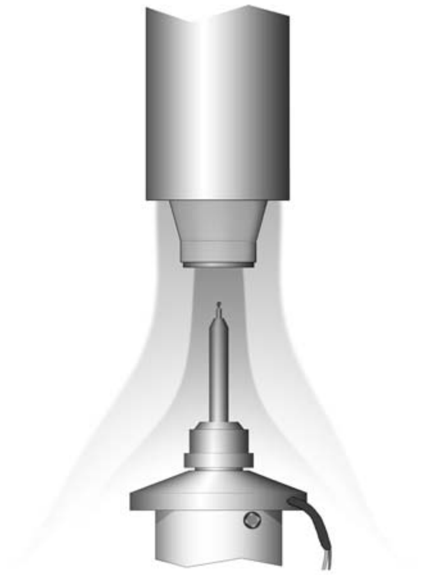


Figure 10.1.4.1
Schematic drawing of a dual-stream setup with the streams parallel to the diffractometer φ axis. The top part represents the outlet end of the stream delivery device. The outer stream (lighter grey) is dry, warm air. The goniometer head is protected by a shield.

For a given setting of the flow control valve, the flow rate depends on the tank pressure. For a constant temperature, a constant pressure is required. The pressure is controlled with a pressure regulator that can reduce the tank pressure by releasing helium gas or raise it by adding helium gas from an external helium supply source. A typical delivery tank pressure is around 20 kPa. It is very important that the pressure is constant. The precise value attained is less important.

At atmospheric pressure, helium at around 40 K has the same density as room-temperature air. This means that very cold helium will rise in air, and has the capacity to seriously cool instrument parts in its way. Goniometer heads, beam stops and beam collimators are particularly vulnerable. A simple remedy is to use a small fan to mix the cold helium with room air.

10.1.3.3. *Frost prevention*

Three areas must be kept frost-free: the crystal, the crystal mount and the delivery end of the transfer tube. The first successful solution to this problem was the dual-stream design of Post *et al.* (1951). It provides for a cold stream surrounded by a concentric warm stream. If the warm stream is sufficiently dry, this will prevent frost around the outlet. The crystal will remain frost-free only if mixing of the two flows occurs downstream from the crystal. For a stream aligned with the axis of the goniometer head, an additional shield is needed to keep the goniometer head frost-free.

10.1.4. Operational considerations

10.1.4.1. *Dual-stream instruments*

Fig. 10.1.4.1 shows a schematic drawing of the region around the crystal in a traditional dual-stream apparatus, first described by Post *et al.* (1951). The device provides for a cold stream surrounded by a concentric warm stream. The diameter of the cold stream is typically around 7 mm with a shield stream of 2–3 mm. The two streams flow parallel to the axis of the crystal mount. In a properly functioning apparatus, the warm stream supplies enough heat to keep the tip of the cold-stream tube above the dew point. It is important that the streams do not mix, or the crystal temperature will not be stable. This is achieved by careful balancing of flow rates to minimize turbulence. (Absence of turbulence can be judged by the shape of the shadow of the cold stream in a parallel beam of bright light.) In a laminar cold stream, the crystal is well protected and no special precautions are needed. The region of constant, minimum temperature will typically have a diameter of about 3 mm. Turbulent flow will result in the absence of any constant-temperature region, so it is vitally important to verify the stream quality.

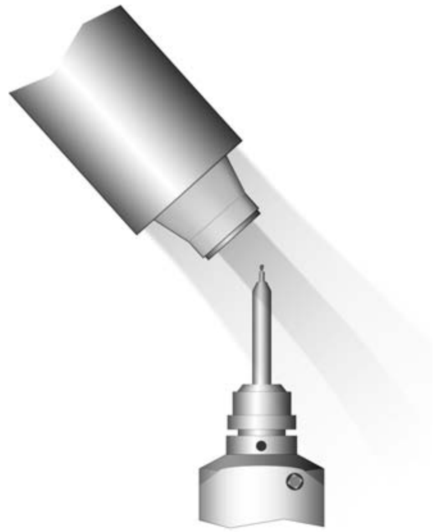


Figure 10.1.4.2
Schematic drawing of a dual-stream setup with the streams angled relative to the diffractometer φ axis. Stream representations are the same as in Fig. 10.1.4.1. The cold stream misses the goniometer head, so no shield is required.

The cold stream has sufficient heat capacity to cool down the goniometer head, and sometimes other adjacent equipment parts as well. A simple solution consists of an aluminium cone equipped with a heating coil on the back. A shield that functions well has been described by Bellamy *et al.* (1994).

Fig. 10.1.4.2 illustrates a situation where the stream direction deviates substantially from the head-on direction in Fig. 10.1.4.1. An angle of 35–55° will give good results. An advantage of an angled delivery is that the cold stream will not touch the goniometer head, and therefore the heated stream deflector is not needed, resulting in simplified installation and operation.

Analysis of the dual-stream apparatus reveals a twofold function of the outer stream: it keeps the nozzle frost-free and it supplies heat to the mounting pin. Protection of the crystal is, in reality, already provided by the laminar cold stream. The nozzle can be kept frost-free simply with an electric heater. Ice formation on the crystal mount can be easily suppressed by appropriate design of the mounting pin and mounting fibre, and attention to their interaction with the cold stream. A successful solution is sketched in Figs. 10.1.4.3 and 10.1.4.4.

10.1.4.2. *Electrically heated nozzle*

Fig. 10.1.4.3 shows the functional equivalent of Fig. 10.1.4.1. Instead of the warm stream, an electrical heating element is used to keep the tip of the delivery tube ice free. An actual construction will usually consist of a nozzle that can be attached to the delivery tube. The heating element is made from standard resistance wire (*e.g.* Nichrome). About 5 W will usually be

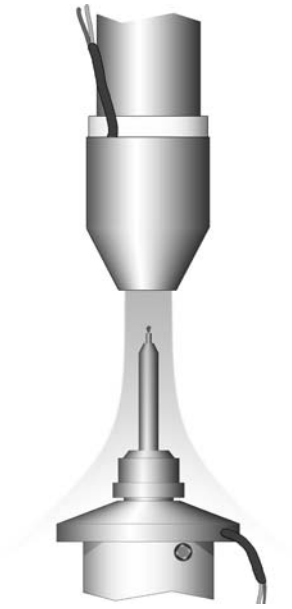


Figure 10.1.4.3
Schematic drawing of a single-stream setup with the stream parallel to the diffractometer φ axis. The cold stream is represented by a grey region. The nozzle is heated above the dew point with a heating coil. The goniometer head is protected by a shield.

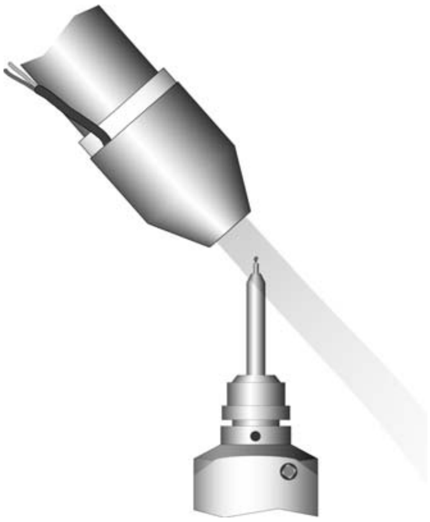


Figure 10.1.4.4
Schematic drawing of a single-stream setup with the stream angled relative to the diffractometer φ axis. The cold stream is represented by a grey region. The crystal mounting pin protrudes 1–2 mm into the cold stream. This prevents frost from forming on the mounting fibre. The nozzle is heated above the dew point with a heating coil. The cold stream misses the goniometer head, so no shield is required. In general, the simplest operation is attained with a setup similar to that shown here.

enough to prevent frost or condensation. The head-on direction results in reliable frost protection for the crystal, but requires that the goniometer head is equipped with a heated stream deflector.

Fig. 10.1.4.4 shows the functional equivalent of Fig. 10.1.4.2. This arrangement leads to the simplest design, although some precautions are needed. The mounting pin itself supplies the heat needed to prevent ice formation on the pin. The tip of the pin should be smooth in order to prevent turbulence. About 1–2 mm (but not more) of the tip must protrude into the cold stream. If too much of the pin is in the cold stream, the rest of the pin can become too cold and ice up. If the pin is too far out of the stream frost prevention will also fail, because glass or other insulating mounting materials will invariably collect ice at the cold/warm interface. As frost prevention depends on heat conducted from the rest of the pin, it must be made from copper (a requirement not strictly necessary for the dual-stream design). This design has been extensively tested and has been used for many years for the collection of a large number of data sets. Results are uniformly good, with simple operation and reliable frost prevention even in high humidity.

Fig. 10.1.4.5 is a photograph of a laboratory implementation of the setup of Fig. 10.1.4.4, with helium as the coolant. Operation is as simple as with nitrogen cooling.

Omission of the warm stream results in significant design simplification. The entire apparatus for production of the outer stream is left out, resulting in real savings in manufacture, operation and maintenance. There is no obvious disadvantage, as ice protection is as good as with the dual stream. The main cost to the user is in the requirement that the mounting system be constructed within somewhat narrower limits, including the requirement of a copper-shafted mounting pin. The fact that operator errors tend to become apparent through frost formation can actually be an advantage. With a dual-stream device, an improperly positioned cold stream or an improperly prepared crystal mount may not produce overt signs, even though the crystal temperature is ill-defined.

In all configurations shown, correct positioning of the cold stream is essential. The centre of the stream should not miss the centre of the diffractometer (and hence the crystal) by more than 0.5 mm.

With helium, the single-stream technique represents the best solution and reliable frost prevention at temperatures down to around 5 K is easily attained. Crystal mounting is simple and the sample is always visible, simplifying centring. In the authors' experience, liquid-helium cooling is as simple as liquid-nitrogen cooling.

10.1.4.3. Temperature calibration

Measurement of the temperature at the crystal site with sensing devices that require attached leads is very difficult, mainly because of heat conduction along the leads. It is usually necessary to loop the leads into the delivery nozzle.

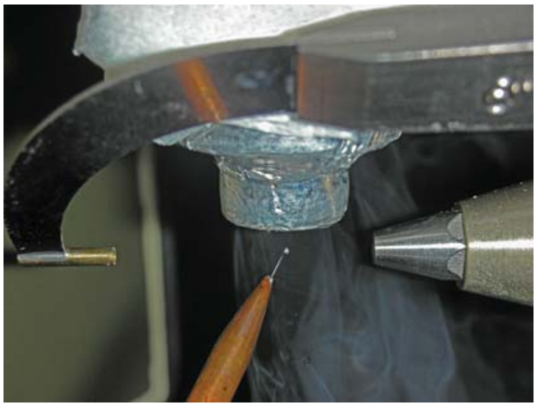


Figure 10.1.4.5
A crystal in a helium stream at 8 K in a setup corresponding to Fig. 10.1.4.4. A thin layer of fog forms at the helium–air interface. The cold stream breaks up well below the crystal position. Rising cold gas (visible as fog to the right of the cold stream) has been mixed with air to prevent cooling of diffractometer parts. The nozzle temperature is 295 K. The crystal is attached to a tapered glass fibre. The light-coloured region at top is not ice; it is part of the insulation. This photograph was taken after the end of data collection.

Both Si diodes and Pt resistance sensors have become sufficiently miniaturized to make them preferred choices at the lowest temperatures. Thermocouples are acceptable above about 80 K. A reliable method of calibration makes use of the known temperature of a phase transition of a crystal in the normal data-collection position. For example, KH_2PO_4 (often referred to as KDP) has a sharp transition at 123 K from tetragonal to orthorhombic, and is commonly used. Another possibility is KH_2AsO_4, which has a corresponding phase transition at 95 K.

Two readout temperatures suffice, one at room temperature and one at the phase transition. The difference between readout temperature and crystal-site temperature can be assumed to vary linearly with T, so interpolation or extrapolation is simple.

10.1.4.4. Transfer of the crystal to the diffractometer

Inspection of Figs. 10.1.4.1–10.1.4.4 reveals that the mounting of a crystal on a mounting pin *via* the traditional placement of the pin in the hole of a standard goniometer head is not simple, because the cooling nozzle is in the way. The solution to the problem is a design that allows side entry. This is most commonly achieved with a magnetic platform on the goniometer head and a corresponding magnetic base on the mounting pin, but an alternative means of side entry employs a slot on a modified goniometer head; the slot is equipped with a spring-loaded catch that allows a very smooth, but stable, catch of the pin.

The use of liquid-N_2 cooling and side entry, and the requirement of reproducible knowledge of crystal temperature at all times, led to the development of a set of tools for crystal mounting as described by Parkin & Hope (1998). The tools include special transfer tongs used for moving crystals from liquid N_2 to the goniometer head. The temperature of the crystal is maintained by the heat capacity and low heat conductance of the tongs. The operation is independent of the orientation of the goniometer head because there is no liquid to contain.

10.1.4.5. Automated robotic crystal handling

The era of structural genomics has necessitated a move to high throughput at high-intensity synchrotron sources. To meet this goal, a number of robotic crystal-handling devices have been developed. These in turn have had the beneficial effect of increased standardization in mounting-pin geometry.

10.1.5. Concluding note

With correctly functioning low-temperature equipment and appropriate techniques, a crystal can be maintained frost-free for the duration of a data-collection run. Formation of frost on the crystal indicates malfunction of the equipment, or operator error. The most likely cause is operator error, but faulty equipment cannot be ruled out. The techniques described here have been used for collecting thousands of data sets from ice-free crystals and crystal mounts. There is no reason to accept frost problems as an unavoidable part of cryocrystallography.

References

Bald, W. B. (1984). *The relative efficiency of cryogenic fluids used in the rapid quench cooling of biological samples. J. Microsc.* **134**, 261–270.

Bellamy, H. D., Phizackerley, R. P., Soltis, S. M. & Hope, H. (1994). *An open-flow cryogenic cooler for single-crystal diffraction experiments. J. Appl. Cryst.* **27**, 967–970.

Brinkmann, C., Weiss, M. S. & Weckert, E. (2006). *The structure of the hexagonal crystal form of hen egg-white lysozyme. Acta Cryst.* D**62**, 349–355.

Chinte, U., Shah, B., Chen, Y.-S., Pinkerton, A. A., Schall, C. A. & Hanson, B. L. (2007). *Cryogenic (<20 K) helium cooling mitigates radiation damage to protein crystals. Acta Cryst.* D**63**, 486–492.

Corbett, M. C., Latimer, M. J., Poulos, T. L., Sevrioukova, I. F., Hodgson, K. O. & Hedman, B. (2007). *Photoreduction of the active site of the metalloprotein putidaredoxin by synchrotron radiation. Acta Cryst.* D**63**, 951–960.

Douzou, P., Hui Bon Hoa, G. & Petsko, G. A. (1975). *Protein crystallography at sub-zero temperatures: lysozyme–substrate complexes in cooled mixed solutions. J. Mol. Biol.* **96**, 367–380.

Esnouf, R. M., Ren, J., Garman, E. F., Somers, D. O'N., Ross, C. K., Jones, E. Y., Stammers, D. K. & Stuart, D. I. (1998). *Continuous and discontinuous changes in the unit cell of HIV-1 reverse transcriptase crystals on dehydration. Acta Cryst.* D**54**, 938–953.

Fu, Z.-Q., Du Bois, G. C., Song, S. P., Harrison, R. W. & Weber, I. T. (1999). *Improving the diffraction quality of MTCP-1 crystals by post-crystallization soaking. Acta Cryst.* D**55**, 5–7.

Garman, E. F. & Mitchell, E. P. (1996). *Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions. J. Appl. Cryst.* **29**, 584–587.

Garman, E. F. & Nave, C. (2009). *Synchrotron radiation damage in protein crystals examined under various conditions by different methods. J. Synchrotron Rad.* **16**, 129–132.

Götz, G., Mészáros, E. & Vali, G. (1991). *Atmospheric particles and nuclei*, p. 142. Budapest: Akadémiai Kiadó.

Haas, D. J. & Rossmann, M. G. (1970). *Crystallographic studies on lactate dehydrogenase at −75 °C. Acta Cryst.* B**26**, 998–1004.

Harp, J. M., Timm, D. E. & Bunick, G. J. (1998). *Macromolecular crystal annealing: overcoming increased mosaicity associated with cryocrystallography. Acta Cryst.* D**54**, 622–628.

Holton, J. M. (2009). *A beginner's guide to radiation damage. J. Synchrotron Rad.* **16**, 133–142.

Hope, H. (1988). *Cryocrystallography of biological macromolecules: a generally applicable method. Acta Cryst.* B**44**, 22–26.

Hope, H. (1990). *Cryocrystallography of biological macromolecules at ultra-low temperature. Annu. Rev. Biophys. Biophys. Chem.* **19**, 107–126.

Hope, H., Frolow, F., von Böhlen, K., Makowski, I., Kratky, C., Halfon, Y., Danz, H., Webster, P., Bartels, K. S., Wittmann, H. G. & Yonath, A. (1989). *Cryocrystallography of ribosomal particles. Acta Cryst.* B**45**, 190–199.

Hui Bon Hoa, G. & Douzou, P. (1973). *Ionic strength and protein activity of supercooled solutions used in experiments with enzyme systems. J. Biol. Chem.* **248**, 4649–4654.

Juers, D. H. & Matthews, B. W. (2004). *The role of solvent transport in cryo-annealing of macromolecular crystals. Acta Cryst.* D**60**, 412–421.

Kriminski, S., Caylor, C. L., Nonato, M. C., Finkelstein, K. D. & Thorne, R. E. (2002). *Flash-cooling and annealing of protein crystals. Acta Cryst.* D**58**, 459–471.

Leidenfrost, J. G. (1756). *De aquae communis nonnullis qualitatibus tractatus*. Duisburg, Germany.

Meents, A., Wagner, A., Schneider, R., Pradervand, C., Pohl, E. & Schulze-Briese, C. (2007). *Reduction of X-ray-induced radiation damage of macromolecular crystals by data collection at 15 K: a systematic study. Acta Cryst.* D**63**, 302–309.

Mitchell, E. P. & Garman, E. F. (1994). *Flash freezing of protein crystals: investigation of mosaic spread and diffraction limit with variation of cryoprotectant concentration. J. Appl. Cryst.* **27**, 1070–1074.

Parkin, S. & Hope, H. (1998). *Macromolecular cryocrystallography: cooling, mounting, storage and transportation of crystals. J. Appl. Cryst.* **31**, 945–953.

Parkin, S. & Hope, H. (2003). *Low-temperature water reconstruction in concanavalin A, with implications for controlled protein crystal annealing. Acta Cryst.* D**59**, 2228–2236.

Petrova, T., Ginell, S., Mitschler, A., Hazemann, I., Schneider, T., Cousido, A., Lunin, V. Y., Joachimiak, A. & Podjarny, A. (2006). *Ultrahigh-resolution study of protein atomic displacement parameters at cryotemperatures obtained with a helium cryostat. Acta Cryst.* D**62**, 1535–1544.

Petsko, G. A. (1975). *Protein crystallography at sub-zero temperatures: cryoprotective mother liquors for protein crystals. J. Mol. Biol.* **96**, 381–392.

Post, B., Schwartz, R. S. & Fankuchen, I. (1951). *An improved device for X-ray diffraction studies at low temperatures. Rev. Sci. Instrum.* **22**, 218–220.

Ryan, K. P. (1992). *Cryofixation of tissues for electron microscopy: a review of plunge cooling methods. Scanning Microsc.* **6**, 715–743.

Schreuder, H. A., Groendijk, H., van der Lan, J. M. & Wierenga, R. K. (1988). *The transfer of protein crystals from their original mother liquor to a solution with a completely different precipitant. J. Appl. Cryst.* **21**, 426–429.

Teng, T.-Y. (1990). *Mounting of crystals for macromolecular crystallography in a free-standing thin film. J. Appl. Cryst.* **23**, 387–391.

Travers, F. & Douzou, P. (1970). *Dielectric constants of alcoholic-water mixtures at low temperature. J. Phys. Chem.* **74**, 2243–2244.

Vali, G. (1995). In *Biological Ice Nucleation and its Applications*, edited by R. E. Lee Jr, G. J. Warren & L. V. Gusta, pp. 1–28. St Paul: APS Press.

Walker, L. J., Moreno, P. O. & Hope, H. (1998). *Cryocrystallography: effect of cooling medium on sample cooling rate. J. Appl. Cryst.* **31**, 954–956.

Weik, M., Kryger, G., Schreurs, A. M. M., Bouma, B., Silman, I., Sussman, J. L., Gros, P. & Kroon, J. (2001). *Solvent behaviour in flash-cooled protein crystals at cryogenic temperatures*. *Acta Cryst.* D**57**, 566–573.

Weik, M., Schreurs, A. M. M., Leiros, H.-K. S., Zaccai, G., Ravelli, R. B. G. & Gros, P. (2005). *Supercooled liquid-like solvent in trypsin crystals: implications for crystal annealing and temperature-controlled X-ray radiation damage studies*. *J. Synchrotron Rad.* **12**, 310–317.

Yeh, J. I. & Hol, W. G. J. (1998). *A flash-annealing technique to improve diffraction limits and lower mosaicity in crystals of glycerol kinase*. *Acta Cryst.* D**54**, 479–480.

International Tables for Crystallography (2012). Vol. F, Chapter 10.2, pp. 249–255.

Chapter 10.2. Cryocrystallography techniques and devices

D. W. Rodgers

10.2.1. Introduction

The ability to collect X-ray data from macromolecular crystals at cryogenic temperatures has played a key role in the more widespread and effective use of crystallography as a tool for biological research. Radiation-damage rates are greatly reduced at low temperature, often by orders of magnitude, making it possible to work with crystals that would otherwise diffract too weakly or decay too quickly. In particular, this radiation protection is essential for the full use of increasingly powerful synchrotron-radiation sources, and the coupling of the two permits the investigation of ever larger macromolecular complexes and the collection of higher-resolution data from nearly all samples. Cryogenic data collection has also allowed efficient experimental phasing using multiple-wavelength methods, as well as preservation of unstable samples and trapping of transient enzyme intermediates, an area that will undoubtedly continue to gain in importance.

In this chapter, practical aspects of cryocrystallography are discussed, with an emphasis on techniques and devices for crystal preparation and handling. A number of prior reviews covering these and other aspects of macromolecular cryocrystallography are available (Hope, 1988, 1990; Watenpaugh, 1991; Rodgers, 1994, 1997; Abdel-Meguid *et al.*, 1996; Garman & Schneider, 1997; Parkin & Hope, 1998). Radiation-damage protection at low temperature, which is not well understood, has also been discussed (Henderson, 1990; Gonzalez & Nave, 1994; Rodgers, 1996; Garman & Schneider, 1997), and the principles and operation of cryostats for sustained cooling during data collection are described elsewhere (Rudman, 1976; Hope, 1990; Garman & Schneider, 1997).

10.2.2. Crystal preparation

Macromolecular crystals are intimately associated with bulk aqueous solution. It surrounds them and penetrates them as solvent-filled channels, which typically account for 30–80% of the crystal volume. A key goal, therefore, of any procedure for cooling these samples to cryogenic temperatures is to prevent the formation of hexagonal crystalline ice. Ice formation, because of the associated increase in specific volume, inevitably disrupts the order of the macromolecular crystals and renders them useless for data collection. The principle that underlies current methods is that sufficiently rapid cooling causes the formation of a rigid glass before ice nucleation can occur. The high viscosity of the glass then prevents subsequent rearrangement into an ordered lattice. Early attempts to flash cool macromolecular crystals were made by Low *et al.* (1966), Haas (1968), and Haas & Rossmann (1970).

With samples as large as macromolecular crystals, however, it is not possible to achieve the high cooling rates necessary to prevent ice formation in water and most aqueous crystallization solutions. The most general method for overcoming this problem is to equilibrate the crystal with a solution containing a cryoprotective agent that slows ice nucleation and allows the formation of a glassy solid with attainable cooling rates. A list of cryoprotectants used successfully with macromolecular crystals is shown in Table 10.2.2.1. Typically, these cryoprotectants are included in the established stabilization or harvest solution at concentrations that range from 6–50%. Glycerol is frequently chosen for initial trials and appears to be a widely applicable cryoprotectant for both salt and organic precipitants. Concentrations of glycerol necessary to prevent ice formation in a number of typical crystallization solutions have been tabulated (Garman & Mitchell, 1996). Other compounds such as ethylene glycol and small sugars fall into this same class. Crystallization precipitants such as 2-methyl-2,4-pentanediol (MPD) or polyethylene glycol (PEG) can often simply be increased in concentration to provide sufficient cryoprotection. Ethanol, methanol and MPD are useful in relatively low-salt conditions. The listed stereoisomer of butanediol is a particularly effective cryoprotectant and can be used where other components of the solution, such as high salt, may limit the amount of cryoprotectant that can be added. Limitations owing to salt in the crystallization mix can also be overcome by transferring the crystals to a solution containing an organic precipitant before or during the introduction of the cryoprotectant (Singh *et al.*, 1980; Ray *et al.*, 1991; Wierenga *et al.*, 1992). Combinations of cryoprotectants have been used where a single cryoprotectant alone did not permit successful flash cooling.

It is rare that a crystal can be transferred without damage directly to a solution containing full-strength cryosolvent. Usually, the cryoprotectant must be introduced slowly to reduce stress on the crystal lattice. Methods of introducing cryoprotectant-containing solutions are listed in Table 10.2.2.2. Techniques such as serial transfer through increasing cryoprotectant concentrations or dialysis are preferred. They allow the crystal to equilibrate with the cryosolvent, leading to reproducible crystal quality and unit-cell dimensions after flash cooling. Also, with equilibrium methods, the solution conditions can be altered in an

Table 10.2.2.1
List of cryoprotectants used successfully for flash cooling macromolecular crystals

See the text as well as Rodgers (1994, 1997), Abdel-Meguid *et al.* (1996), and Garman & Schneider (1997) for additional details.

(2*R*,3*R*)-(−)-Butane-2,3-diol	2-Methyl-2,4-pentanediol
Erythritol	Polyethylene glycol 400
Ethanol	Polyethylene glycol 1000–10 000
Ethylene glycol	Propylene glycol
Glucose	Sucrose
Glycerol	Xylitol
Methanol	

Table 10.2.2.2
Methods for introducing the cryoprotectants needed for flash cooling

(1) Serial transfer into increasing strengths of cryoprotectant
(2) Dialysis
(3) Growth in cryoprotectant
(4) Brief transfer before flash cooling
(5) Direct transfer into full-strength cryoprotectant

Copyright © 2012 International Union of Crystallography

attempt to control any crystal damage associated with the flash-cooling process itself. The best scheme for serial transfer must be determined empirically. Equilibration time at each step depends on a number of factors (size of the crystals, solvent content, viscosity of the solution) but can be as rapid as less than a minute for small cryoprotectants or as long as hours for large polymers (Bishop & Richards, 1968; Fink & Petsko, 1981; Ray *et al.*, 1991). Typically, 5% increments in cryoprotectant concentration with equilibration times of 15 min or longer at each step are used in initial trials. The step size is then decreased if damage occurs. Dialysis can be done conveniently in the small buttons used for crystallization. These are available with chamber sizes as low as 5 µl and allow a piece of dialysis membrane to be stretched and held securely over the opening. Dialysis times range from 1 to 24 h depending on the size of the cryoprotectant and the viscosity of the solution.

Another technique, growth of the crystals directly in cryoprotectant solutions, is particularly convenient and effective. In some cases, the primary precipitant, MPD for example, may provide cryoprotection if the concentration used in crystallization is sufficiently high. More commonly, however, additives such as glycerol are included in the crystallization buffer. An advantage of this technique is that crystals can be mounted directly from the crystallization drop, eliminating potential damage in transferring to a harvest or cryoprotective solution.

When it is not possible to identify a cryosolvent compatible with the crystals, a brief exposure to the cryoprotective solution may allow successful flash cooling. Apparently, the water in crystal solvent channels is constrained sufficiently to prevent nucleation, and simply exchanging the external aqueous solution with cryosolvent provides protection. The 'quick dunk' in the cryosolvent may be as short as a few seconds, and for some crystals it is possible to combine this technique with prior equilibration in lower, non-damaging concentrations of cryoprotectant. The same principle of preventing ice formation in the external solution forms the basis of an alternative technique developed by Hope (1988). Here, the external solution is replaced by a hydrocarbon oil before flash cooling.

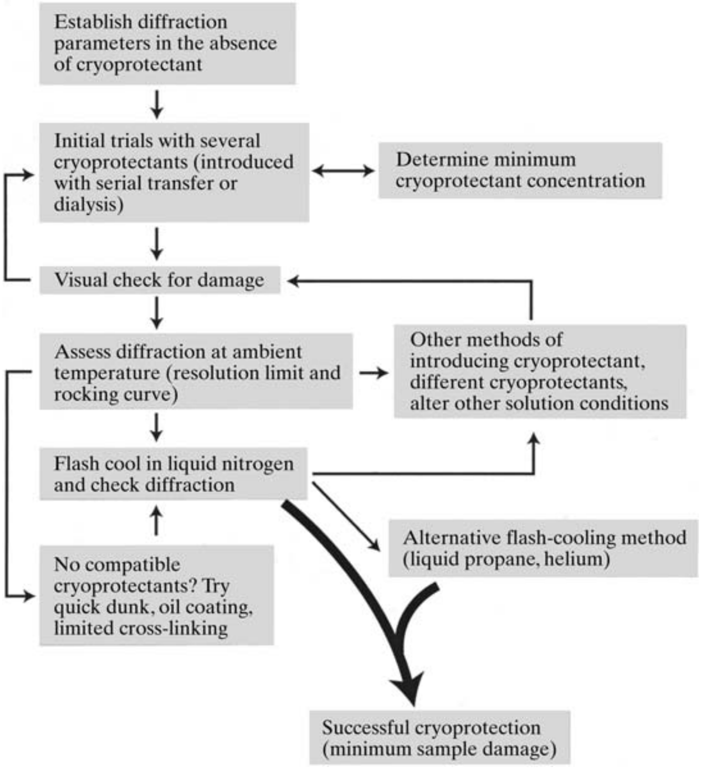

Establish diffraction parameters in the absence of cryoprotectant
Initial trials with several cryoprotectants (introduced with serial transfer or dialysis)
Determine minimum cryoprotectant concentration
Visual check for damage
Assess diffraction at ambient temperature (resolution limit and rocking curve)
Other methods of introducing cryoprotectant, different cryoprotectants, alter other solution conditions
Flash cool in liquid nitrogen and check diffraction
Alternative flash-cooling method (liquid propane, helium)
No compatible cryoprotectants? Try quick dunk, oil coating, limited cross-linking
Successful cryoprotection (minimum sample damage)

Figure 10.2.2.1
Recommended pathway for optimizing cryoprotectant conditions and flash cooling.

Finding suitable cryoprotection conditions is a trial-and-error process. Two problems must be overcome: the cryoprotectant must be introduced without significant damage to the crystal, and damage during the flash-cooling process must be minimized. A scheme for systematically determining conditions for flash cooling is given in Fig. 10.2.2.1. In order to assess the effect of subsequent manipulations, it is important first to establish the resolution and rocking curve of the crystals under normal harvest conditions. Then one or a few cryoprotectants can be added to the harvest solution under conditions that allow equilibration with the crystal. The minimum concentration of cryoprotectant necessary to prevent ice formation can be determined by flash cooling candidate solutions using the loop-mounting technique described in Section 10.2.3. A sufficient concentration of cryoprotectant will result in a transparent glass upon cooling, while too low a concentration will produce opaque microcrystalline ice. A solution of cryoprotectant 2–3% above this minimum value should be used to allow for the added volume and therefore slower cooling when the crystal is present. If the crystals crack or dissolve in a cryosolvent, then the cryoprotectant should be introduced more slowly, the solution conditions (precipitant concentration, ionic strength, pH) altered, or the cryoprotectant eliminated from consideration.

The diffraction quality of crystals that show no visible sign of damage should be assessed at the crystal-growth temperature, and solution conditions should be altered if there has been a significant loss in resolution or an increase in rocking-curve width. For crystals that are incompatible with a wide range of cryosolvent conditions, quick-dunk and oil-coating techniques should be considered. Limited cross-linking (with glutaraldehyde, for example) can sometimes stabilize crystals for the introduction of cryoprotectant or improve stability during flash cooling.

When conditions that result in little or no damage have been identified, the crystals should be flash cooled and the diffraction assessed again. The formation of even small amounts of microcrystalline ice can be detected after flash cooling as characteristic powder rings at low-order spacings of 3.90, 3.67, 3.44 Å. If ice forms, a greater concentration of cryoprotectant must be used. An increase in the rocking-curve width of the crystal at this stage is common, probably due to the thermal stress on the lattice or changes in solution properties on cooling. If this increase is more than 50%, or if any loss of resolution occurs, solution conditions should be altered and the process repeated. The concentration of the cryoprotectant can be increased and different cryoprotectants tested. Other solution parameters, as noted above, can also be adjusted in an attempt to decrease the damage from flash cooling. In addition, different flash-cooling techniques (discussed below) can be tested to determine whether they produce less damage. Suitable cryosolvent conditions are usually established after a few trials, and even in difficult cases it has generally proven possible to find acceptable conditions by continuing to refine solution parameters.

10.2.3. Crystal mounting

A mounting technique suitable for flash cooling should allow for rapid heat exchange by providing a large surface area and a minimum of extraneous material that must be cooled. The technique should also subject the crystals to little mechanical

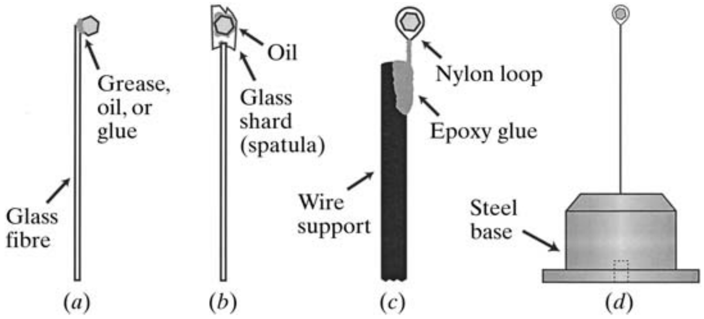


Figure 10.2.3.1
Different crystal mounts for flash cooling and cryogenic data collection. (*a*) Crystal mounted on a thin glass fibre with adhesive, grease, or oil. (*b*) Crystal placed in a hydrocarbon oil and then scooped onto a thin glass shard. (*c*) Crystal suspended in a film of aqueous solution within a nylon loop. The loop is attached to a thin (~0.25 mm diameter) wire support. (*d*) A diagram of the entire loop-mount assembly. The base is made of plain steel or a magnetic alloy and has two holes, one for the wire post and one for a locating pin, which reproducibly positions the assembly on the goniometer.

stress and should result in a relatively compact sample that can be immersed in the narrow gas stream used to maintain the temperature during data collection. The glass capillary tubes conventionally used to mount macromolecular crystals are not well suited to flash-cooling procedures since they insulate the sample, reducing cooling rates, and their bulk interferes with cryogenic equipment. A number of alternative mounting methods used for flash cooling are shown in Fig. 10.2.3.1. Crystals can be affixed directly to thin glass fibres with cement or grease (Haas & Rossmann, 1970; Dewan & Tilton, 1987), or they can be scooped up on thin glass spatulas, a procedure first used in conjunction with the oil-coating method described in Section 10.2.2 (Hope, 1988). A loop-mounting technique introduced by Teng (1990) has proven the most generally applicable, however, and has become the method of choice. Here, the crystal is held suspended in a thin film of cryosolvent formed in a small loop. The technique is quick and straightforward, remarkably gentle to the crystal, and provides a large surface area for cooling.

The loops are generally formed from nylon fibre, although glass wool is useful for larger versions because its rigidity keeps them from collapsing under the surface tension of the suspended film. Both types of fibres should have a diameter of approximately 10 µm. This small cross section reduces absorption and scattering from the material itself and also minimizes the thickness of the film in the loop. Several methods of making the loops have been described in detail (Rodgers, 1997; Garman & Schneider, 1997), and nylon loops of different sizes are available commercially. The loop is usually glued to a thin metal wire or other heat-conductive post. The ability to conduct heat rapidly is required to minimize ice formation at the point where the wire or post exits the cold gas stream of the cryostat, which occurs in some orientations of the loop assembly. This post is in turn attached to a steel base, which is used with the magnetic transfer system described below.

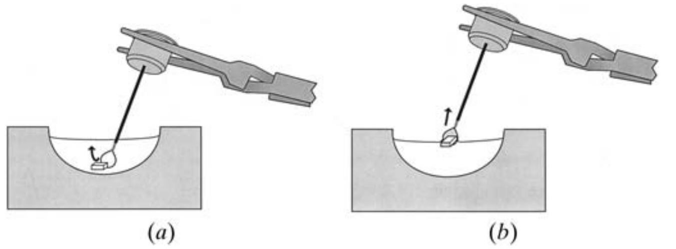


Figure 10.2.3.2
Mounting a crystal in a loop. (*a*) While viewing with a stereomicroscope, the crystal is teased to the surface of the liquid using the loop. (*b*) It is then drawn through the interface and into the loop. The sizes of the loop and crystal have been exaggerated. Reproduced with permission from Rodgers (1997). Copyright (1997) Academic Press.

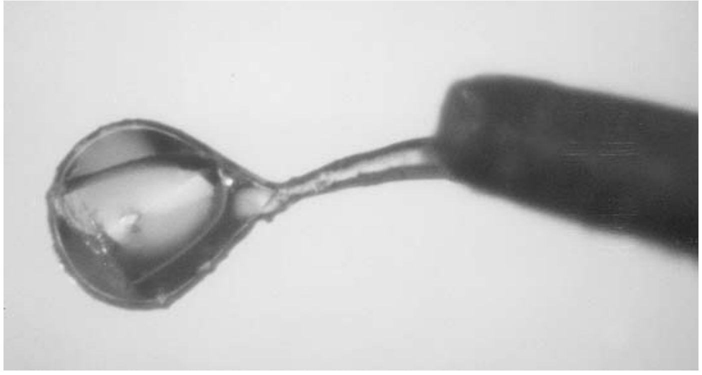


Figure 10.2.3.3
Photograph of a flash-cooled crystal mounted in a nylon loop. The wire post holding the loop is visible on the right. Reprinted from Rodgers (1994) with permission from Elsevier Science.

Crystals are placed in the loop as shown in Fig. 10.2.3.2. They can be mounted directly from the crystallization drop or after harvesting into any convenient container. Under a stereomicroscope, the crystal is teased to the surface of the solution, usually with the loop itself. Once at the surface, the crystal is carried through the interface by first resting it on the bottom of the loop and then moving the assembly vertically to pull it out of the solution. A practiced experimentalist can usually capture the crystal in the first few tries. The plane of the loop should be kept near the vertical to increase the chance of catching the crystal and to minimize the amount of liquid drawn up with it. An alternative technique is to use a small pipette to place the crystal and a drop of cryosolvent into the loop and then draw off the excess solution with filter paper. In either case, it can be difficult to form a film in

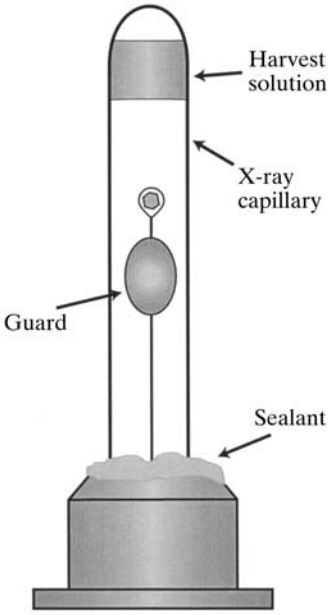


Figure 10.2.3.4
Arrangement for using the loop-mounting technique at non-cryogenic temperatures.

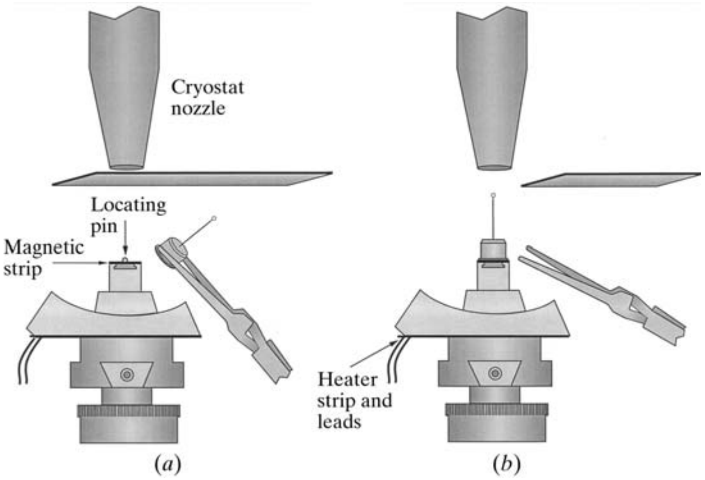


Figure 10.2.4.1
Flash cooling a crystal in the cold gas stream from a cryostat. (*a*) The cold gas from the nozzle is blocked and the loop assembly placed on the goniometer. A locating pin on the goniometer ensures reproducible positioning of the loop assembly, which is held in place by a magnetic strip. (*b*) The gas stream is then unblocked to rapidly cool the crystal. A heating element in contact with the goniometer keeps it from icing during data collection. Based on a diagram by Rodgers (1997).

the loop with solutions high in organic solvent due to the lack of surface tension. For these solutions, adding PEG up to a few per cent usually allows a stable film to form. Fig. 10.2.3.3 is a photograph of a crystal mounted in a nylon loop. If the diameter of the loop is chosen so that it just accommodates the crystal, mounting is easier and the amount of extra scattering material in the X-ray beam is reduced. Also, asymmetric crystals can then be oriented relative to the assembly by preforming the loop into the appropriate shape.

The loop-mounting technique can also be used for data collection above cryogenic temperatures by sealing the loop and pin in a large diameter (3 mm) glass or quartz X-ray capillary (Fig. 10.2.3.4). A guard composed of stiff wax or a plastic plug cemented to the pin helps to guide the capillary over the sample before sealing it to the base with high vacuum grease or a cement low in volatile solvent. Loop mounting can be less damaging for many crystals than capillary mounting, and it results in a more uniform X-ray absorption surface.

10.2.4. Flash cooling

Once mounted in the loop, the crystal must be cooled rapidly to prevent ice formation. A simple and often effective approach (see Hope, 1990; Teng, 1990) is to flash cool the sample in a cryostat gas stream (most frequently nitrogen, but also helium) right on the X-ray camera. This technique has the added advantage of leaving the crystal in position for immediate analysis and data collection. As shown in Fig. 10.2.4.1, the gas stream from the cryostat nozzle is temporarily deflected while the loop assembly is placed on the goniometer of the X-ray camera. The stream is then unblocked, allowing the cold gas to flow over the crystal. Deflecting the cold stream before placing the loop assembly eliminates the risk that the sample will cool slowly and form ice in the warmer outer layers of the gas stream. The arrangement of the cryostat nozzle shown in Fig. 10.2.4.1, with the gas stream coaxial to the loop assembly, is particularly effective. The cooling gas (usually at around 110 K for nitrogen cryostats) flows across both surfaces of the loop, maximizing the rate and evenness of cooling. Other orientations of the nozzle are frequently used, and in those cases the loop should be aligned with one edge pointing at the incoming gas. Note that a heating element, as shown in Fig. 10.2.4.1, is required to prevent icing of the goniometer with the nozzle in the coaxial position.

When handling the loop-mounted crystal before flash cooling, care must be taken to avoid drying the sample. The same characteristics that make the loop mount so effective for flash cooling, a large surface area and a small amount of surrounding solution, also promote a rapid loss of water and any other volatile component. The resulting change in solute concentration can damage the crystal or result in non-isomorphism between crystals. For this reason, every effort should be made to reduce the time required to flash cool the crystal after it is mounted. One key to avoiding delay when flash cooling in the cold stream is a rapid and reliable method of attaching the loop assembly to the goniometer. A magnetic mounting system (Fig. 10.2.4.1) developed by Rodgers (1994, 1997) is frequently used. Here, either a portion of flexible magnetic strip or solid magnet is affixed to the goniometer to hold the ferromagnetic base of the loop assembly. The base is positioned reproducibly by a small locating pin protruding from the goniometer, which mates with the centred hole in the loop base (Fig. 10.2.3.1*d*). A second pin or key can be used to specify the orientation of the loop assembly about its axis if necessary.

While flash cooling in the cold stream is convenient, an alternative method, rapidly plunging the crystal into a liquid cryogen, offers several advantages. This technique generally results in more even cooling of both sides of the loop-mounted sample, which may decrease damage due to thermal stress (Haas & Rossmann, 1970). It also reduces the time between mounting the crystal and flash cooling, and it can be used easily in any location – a cold room, for example. Another possible advantage of the liquid-cryogen method is that it produces a higher cooling rate than the cryostat gas stream, at least over much of the temperature range traversed during cooling (Walker *et al.*, 1998; Teng & Moffat, 1998). With increased cooling rates, the percentage of cryoprotectant necessary to prevent ice formation is lower, an advantage when benign cryoprotectant conditions prove difficult to find. Changes in solution dielectric or other parameters may also cause less damage. On the other hand, although cooling may be more even in a liquid cryogen, the overall increase in cooling rate could result in even greater thermal gradients, and therefore greater thermal stress, across the crystal. Systematic studies are needed to assess the effect of cooling rate on the quality of flash-cooled crystals, but in practice the liquid-cryogen technique has proven effective and is widely used.

Common cryogens for flash cooling are liquid nitrogen, propane, and, to a lesser extent, ethane and some types of Freon. (Another potentially useful cryogen, liquid helium, has not yet been explored for flash cooling macromolecular crystals.) There is some disagreement about relative cooling rates in liquid nitrogen *versus* liquid propane for samples the size of loop-mounted crystals (Walker *et al.*, 1998; Teng & Moffat, 1998), but both cryogens are known to work well for flash cooling. Since liquid nitrogen is simpler to use and safer than propane, it should be considered for initial trials with a new type of crystal. A diagram showing flash cooling with liquid nitrogen is presented in Fig. 10.2.4.2(*a*). The crystal is captured in the loop and quickly plunged into a Dewar filled with liquid nitrogen. Attaching the loop assembly to a short rod equipped with a magnetic mount allows it to be plunged deeply into the liquid nitrogen, which may increase the cooling rate by preventing the build-up of insulating

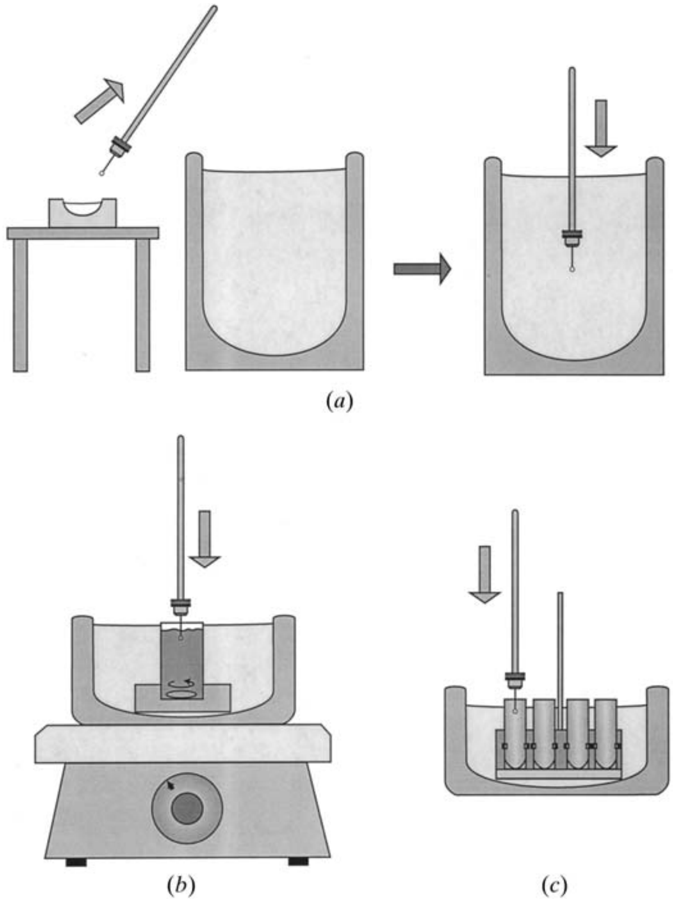


Figure 10.2.4.2
Flash cooling in a liquid cryogen. (*a*) Cooling in liquid nitrogen. The loop assembly is attached, *via* a magnet mount, to a short rod and the crystal is captured. It is then quickly plunged into a nearby Dewar of liquid nitrogen. (*b*) One method of flash cooling in a liquid cryogen such as propane. The cryogen is placed in a weighted container, which itself stands in a Dewar of liquid nitrogen. The Dewar rests on a stir plate, which mixes the liquid cryogen to ensure a uniform temperature. When the temperature of the cryogen is just above its melting point, the loop assembly is plunged into the liquid. (*c*) A variation on cooling in propane or a similar cryogen. The cryogen is placed into small plastic vials designed for cryogenic storage. Just before the cryogen freezes, the loop assembly is plunged directly into a vial. A holder for the vials allows multiple samples to be prepared sequentially.

gas around the crystal. To minimize drying of the sample during transfer, the crystal container and the Dewar are located as close as possible. If necessary, drying can be further reduced by using a portable humidifier to add moisture in the area.

Other cryogens, such as propane, can be tested if results with liquid nitrogen are not satisfactory. Two methods for flash cooling in these other cryogens are illustrated in Figs. 10.2.4.2(*b*) and (*c*). In the first (Fig. 10.2.4.2*b*), the liquid cryogen is held in a small container with a weighted base, which is placed in a Dewar of liquid nitrogen to cool the cryogen. The cryogenic liquid is mixed using a magnetic stir bar to ensure a uniform temperature throughout the sample. Since the boiling points of these cryogens are well above their melting points, it is possible in the absence of stirring to have relatively warm, and therefore less effective, cryogen near the top of the container. When a temperature probe indicates that the cryogen is just above its melting point, the crystal is mounted and plunged quickly into the liquid. A variant of this technique (Fig. 10.2.4.2*c*) calls for plunging the loop assembly directly into cryogen-filled plastic vials, which are used for low-temperature transfer and storage of the crystals (see Section 10.2.5). The cryogen is then allowed to solidify around the crystal before it is placed on the X-ray camera or stored for later use. With this technique, it is more difficult to ensure that the temperature of the cryogen is uniform throughout the container. Other mechanisms for flash cooling in liquid cryogens have been described (Hope *et al.*, 1989; Abdel-Meguid *et al.*, 1996), and devices for combining xenon derivatization with flash cooling (Soltis *et al.*, 1997) are available commercially.

10.2.5. Transfer and storage

Crystals flash cooled in a liquid cryogen must be placed for data collection in the cold gas stream of a cryostat without any substantial warming. One common transfer method (Rodgers, 1994, 1997) is shown in Fig. 10.2.5.1. Once the loop assembly has been plunged into the Dewar of liquid nitrogen, it is inserted into a small plastic vial of the type normally used for cryogenic storage, ensuring that the sample remains below the liquid surface during the operation. There is then sufficient liquid nitrogen in the vial to keep the sample cold as it is transferred to the cryostat gas stream. Again, the magnetic mounting system is used to reduce the time required for transfer. For X-ray cameras with vertical spindles, as shown in Fig. 10.2.5.1, some means of pointing the magnetic mount downward is required to prevent the nitrogen from spilling out of the vial. The goniometer illustrated has a detachable arc extension (Engel *et al.*, 1996; Litt *et al.*, 1998) that provides this capability. When the loop assembly is attached to the magnet, the vial is quickly withdrawn, exposing the crystal to the gas stream. The arc slide can then be returned to the normal position and the arc extension removed.

Another device (Mancia *et al.*, 1995) for achieving the correct transfer geometry is shown in Fig. 10.2.5.2. This ‘flipper’ mechanism can be extended to permit transfer of the crystal. The device is then rotated about the hinge to reorient the loop assembly for data collection. The

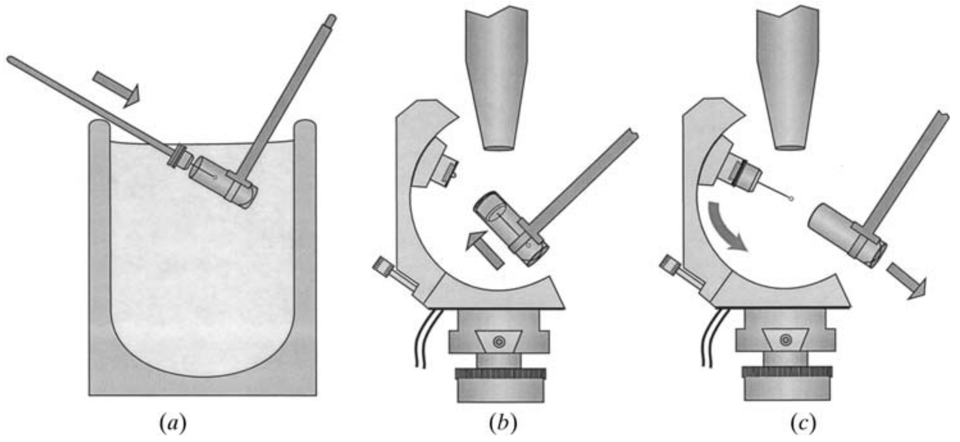


Figure 10.2.5.1
Transfer of the flash-cooled crystal and loop assembly to a goniometer using a cryovial. (*a*) The loop assembly with a flash-frozen crystal is placed in the vial, which is held by a rod-shaped tool. The operation is carried out beneath the surface of the liquid nitrogen in a Dewar. (*b*) The loop assembly is transferred to the goniometer using the magnetic mounting system. (*c*) The vial is withdrawn, exposing the crystal to the cold gas stream. With this arrangement of goniometer and cryosystem nozzle, it is necessary to use a device that allows the magnetic mount to point downward. Here, a detachable arc extension provides this ability. After crystal transfer, the arc slide can be returned to the normal position and the extension removed.

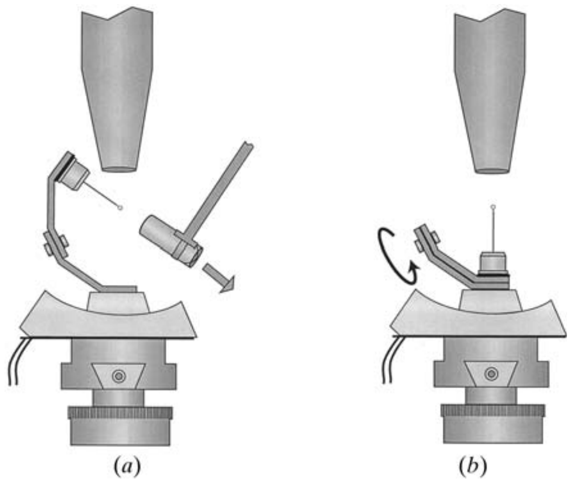


Figure 10.2.5.2
Transfer using an alternative device for achieving the correct magnetic mount orientation. (*a*) A hinged mechanism is extended to orient the magnetic mount downward, and the loop assembly is attached. (*b*) The mechanism is rotated about the hinge to place the mount in the normal orientation for data collection.

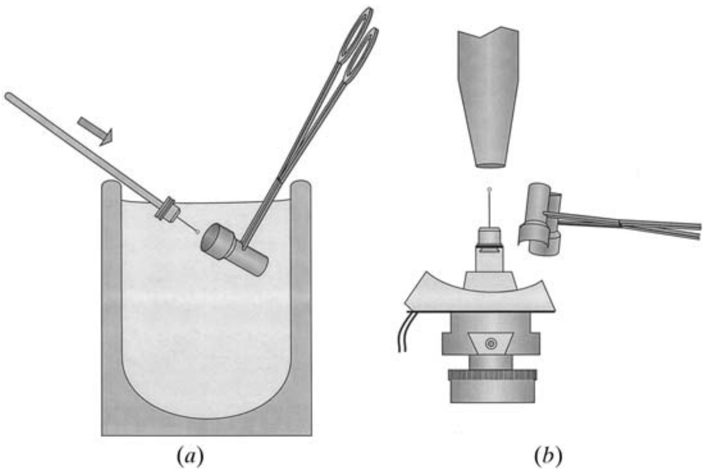


Figure 10.2.5.3
Transfer using tongs. (*a*) The crystal is inserted into a split metal cup that allows the base to be held securely. (*b*) The tongs are inverted and used to place the loop assembly on the magnetic mount. The jaws of the tongs are then opened to separate the halves of the cup, and the tongs are withdrawn.

hinge is positioned so that rotation does not translate the crystal, keeping it in the cold stream during reorientation.

When cooling in other liquid cryogens such as propane, the same cryovial transfer system is used. Flash cooling in cryovials (Fig. 10.2.4.2*c*) permits direct transfer using the magnetic mounting system. Usually, the liquid cryogen has been solidified in the vial, and it is allowed to melt at least partially before placing the crystal on the goniometer. Any remaining solid then melts and drips away (although it is often necessary to remove the last drop on the crystal with filter paper). When cooling in a larger volume of cryogen (Fig. 10.2.4.2*b*), the crystal can be 'hopped' rapidly from the cooling cryogen to the surrounding vat of liquid nitrogen. A drop of cryogen transfers with the crystal, keeping it from warming. The loop assembly can then be placed in a cryovial and transferred to the goniometer.

Another device that does not use cryovials has been introduced (Parkin & Hope, 1998) to facilitate transfer from liquid nitrogen. The device consists of a split metal cup attached to handles that allow the cup to be opened and closed. When closed, the two halves of the metal cup form a cavity that can accommodate and grasp the loop assembly. As shown in Fig. 10.2.5.3, the loop assembly is inserted after first cooling the tongs in liquid nitrogen. The thermal mass of the tongs prevents warming as the crystal is then placed on the goniometer. The tongs are opened and removed to expose the crystal to the gas stream.

Any of these transfer procedures can be reversed in order to return the loop assembly to liquid nitrogen without thawing the crystal. The assembly and cryovial can then be placed in a Dewar designed for long-term storage. Some opening should be present in the loop-assembly bases, or the cryovials should be notched, to allow free movement of liquid nitrogen. The vials are conveniently held and organized using aluminium canes, which take up to five samples and have tabs that hold the loop assemblies in place. For even more secure long-term storage, loop assemblies with threaded bases that screw into the cryovials are available.

The ability to store samples for long periods of time permits a number of crystals to be flash cooled under consistent conditions, which can be important for maintaining isomorphism, and crystals can also be stockpiled for later data collection at a synchrotron X-ray facility. In fact, crystals should be prescreened for quality in the laboratory before synchrotron data collection to make efficient use of time on the beam line. Finally, crystals that degrade in growth or harvest solutions, or that contain macromolecules in unstable or transient states, can be conveniently preserved by flash cooling and storage in liquid nitrogen.

References

Abdel-Meguid, S. S., Jeruzalmi, D. & Sanderson, M. R. (1996). *Crystallographic Methods and Protocols*, edited by C. Jones, B. Mulloy & M. R. Sanderson, pp. 55–87. New Jersey: Humana Press.

Bishop, W. H. & Richards, F. M. (1968). *Properties of liquids in small pores. Rates of diffusion of some solutes in cross-linked crystals of β-lactoglobin. J. Mol. Biol.* **38**, 315–328.

Dewan, J. C. & Tilton, R. F. (1987). *Greatly reduced radiation damage in ribonuclease crystals mounted on glass fibres. J. Appl. Cryst.* **20**, 130–132.

Engel, C., Wierenga, R. & Tucker, P. A. (1996). *A removable arc for mounting and recovering flash-cooled crystals. J. Appl. Cryst.* **29**, 208–210.

Fink, A. L. & Petsko, G. A. (1981). *X-ray cryoenzymology. Adv. Enzymol. Relat. Areas Mol. Biol.* **52**, 177–246.

Garman, E. F. & Mitchell, E. P. (1996). *Glycerol concentrations required for cryoprotection of 50 typical protein crystallization solutions. J. Appl. Cryst.* **29**, 584–587.

Garman, E. F. & Schneider, T. R. (1997). *Macromolecular cryocrystallography. J. Appl. Cryst.* **30**, 211–237.

Gonzalez, A. & Nave, C. (1994). *Radiation damage in protein crystals at low temperature. Acta Cryst.* D**50**, 874–877.

Haas, D. J. (1968). *X-ray studies on lysozyme crystals at −50 °C. Acta Cryst.* B**24**, 604.

Haas, D. J. & Rossmann, M. G. (1970). *Crystallographic studies on lactate dehydrogenase at −75 °C. Acta Cryst.* B**26**, 998–1004.

Henderson, R. (1990). *Cryoprotection of protein crystals against radiation damage in electron and X-ray diffraction. Proc. R. Soc. London Ser. B*, **241**, 6–8.

Hope, H. (1988). *Cryocrystallography of biological macromolecules: a generally applicable method. Acta Cryst.* B**44**, 22–26.

Hope, H. (1990). *Crystallography of biological macromolecules at ultra-low temperature. Annu. Rev. Biophys. Biophys. Chem.* **19**, 107–126.

Hope, H., Frolow, F., von Böhlen, K., Makowski, I., Kratky, C., Halfon, Y., Danz, H., Webster, P., Bartels, K. S., Wittmann, H. G. & Yonath, A. (1989). *Cryocrystallography of ribosomal particles. Acta Cryst.* B**45**, 190–199.

Litt, A., Arnez, J. G., Klaholz, B. P., Mitschler, A. & Moras, D. (1998). *A eucentric goniometer head sliding on an extended removable arc modified for use in cryocrystallography. J. Appl. Cryst.* **31**, 638–640.

Low, B. W., Chen, C. C. H., Berger, J. E., Singman, L. & Pletcher, J. F. (1966). *Studies of insulin crystals at low temperatures: effects on mosaic character and radiation sensitivity. Proc. Natl Acad. Sci. USA*, **56**, 1746–1750.

Mancia, F., Oubridge, C., Hellon, C., Woollard, T., Groves, J. & Nagai, K. (1995). *A novel device for the recovery of frozen crystals. J. Appl. Cryst.* **28**, 224–225.

Parkin, S. & Hope, H. (1998). *Macromolecular cryocrystallography: cooling, mounting, storage and transportation of crystals. J. Appl. Cryst.* **31**, 945–953.

Ray, W. J. Jr, Bolin, J. T., Puvathingal, J. M., Minor, W., Liu, U. & Muchmore, S. W. (1991). *Removal of salt from a salt-induced protein crystal without cross-linking. Preliminary examination of desalted crystals of phosphoglucomutase by X-ray crystallography at low temperature. Biochemistry*, **30**, 6866–6875.

Rodgers, D. W. (1994). *Cryocrystallography. Structure*, **2**, 1135–1140.

Rodgers, D. W. (1996). *Cryocrystallography of macromolecules. Synchrotron Radiat. News*, **9**, 4–11.

Rodgers, D. W. (1997). *Practical cryocrystallography. Methods Enzymol.* **276**, 183–203.

Rudman, R. (1976). *Low-Temperature X-ray Diffraction.* New York: Plenum Press.

Singh, T. P., Bode, W. & Huber, R. (1980). *Low-temperature protein crystallography. Effect on flexibility, temperature factor, mosaic spread, extension and diffuse scattering in two examples: bovine trypsinogen and Fc fragment. Acta Cryst.* B**36**, 621–627.

Soltis, S. M., Stowell, M. H. B., Wiener, M. C., Phillips, G. N. Jr & Rees, D. C. (1997). *Successful flash-cooling of xenon-derivatized myoglobin crystals. J. Appl. Cryst.* **30**, 190–194.

Teng, T.-Y. (1990). *Mounting of crystals for macromolecular crystallography in a free-standing thin film. J. Appl. Cryst.* **23**, 387–391.

Teng, T.-Y. & Moffat, K. (1998). *Cooling rates during flash cooling. J. Appl. Cryst.* **31**, 252–257.

Walker, L. J., Moreno, P. O. & Hope, H. (1998). *Cryocrystallography: effect of cooling medium on sample cooling rate. J. Appl. Cryst.* **31**, 954–956.

Watenpaugh, K. D. (1991). *Macromolecular crystallography at cryogenic temperatures. Curr. Opin. Struct. Biol.* **1**, 1012–1015.

Wierenga, R. K., Zeelen, J. Ph. & Noble, M. E. M. (1992). *Crystal transfer experiments carried out with crystals of trypanosomal triosephosphate isomerase (TIM). J. Cryst. Growth*, **122**, 231–234.

International Tables for Crystallography (2012). Vol. F, Chapter 10.3, pp. 256–261.

Chapter 10.3. Radiation damage

E. F. Garman

10.3.1. Introduction

Radiation damage to macromolecular crystalline samples during the experiment is a problem inherent in using ionizing radiation to obtain diffraction patterns, and has presented a challenge to protein crystallography (MX) since the beginning of the field, often necessitating the use of many crystals to assemble a complete data set. The damage manifests itself as a decrease in diffraction intensity caused by the destruction of crystallinity. This in turn is a result of the absorption in the sample of photons in the beam by either the photoelectric effect (total absorption of the photon and ejection of an inner-shell electron) or Compton scattering (inelastic scattering of the photon which escapes, following a varying amount of energy loss to an atomic electron). At the incident energies used for MX, the former effect has a much higher cross section and dominates the absorption, accounting for over 90% of the energy deposited by the beam. Each photoelectron has enough energy to subsequently induce up to 500 ionization events, which in turn can result in the formation of radical species in the crystal. In protein crystals, the presence of anything between 20 and 80% solvent means that the radiolysis of water and other components of the solvent is an important contributor to the creation of these species. Some of the energy deposited by the beam by these processes is converted into heat and induces a temperature rise in the sample. The diffracted photons are scattered elastically and thus do not contribute to the damage.

The usage of the terms 'primary', 'secondary' and 'tertiary' damage is dependent on the context. The definitions adopted here are as follows: primary damage is the ionization of an atom due to photoelectric absorption or Compton scattering, whereas secondary damage is the formation of up to 500 lower-energy secondary electrons per primary absorption event. These have a temperature-independent continuous slowing down approximation range and induce further ionization and excitation events. The mobilities of the subsequent carriers of energy are reduced at lower temperatures, while the primary photoelectron has a temperature-independent mean track length of a few micrometres (for 12 keV photons) (O'Neill *et al.*, 2002). Tertiary damage is the effect on the crystal lattice and other mechanical consequences of the energy loss in the sample. Damage can also be classified as direct if the primary absorption event occurs at an atom in the protein molecule, or indirect if the radiation is absorbed by the surrounding solvent and the reactive species formed subsequently interact with the protein.

The universal metric against which the decay indicators of a crystal are conveniently measured is the absorbed dose, defined as the energy absorbed per unit mass of the sample (1 Gy = 1 J kg^{-1}). To calculate this requires knowledge of the sample size and composition (*i.e.* the number of each atom type in the unit cell) so that its absorption coefficients can be computed, and detailed information about the incident beam [energy, size, shape and flux (in photons s^{-1})].

The earliest investigation of radiation damage in MX was carried out at room temperature (RT) by Blake & Phillips (1962), who concluded that the damage was proportional to the absorbed dose and that the observed form of the decay with dose could be described by a first-order exponential function. They deduced that a single 8 keV X-ray photon absorbed by the crystal disrupted around 70 protein molecules, and disordered a further 90 for doses up to about 20 Mrad (0.2 MGy, since 1 Gy = 1 J kg^{-1} = 100 rad), above which the decay of the undamaged part of the crystal was faster than expected from an exponential law, due perhaps to some sort of cumulative effect. They also suggested that specific structural damage was suffered by the protein molecule; these conclusions were reached without knowledge of either the sequence or three-dimensional structure of the protein.

Despite reports from the early days of synchrotron use that crystals had longer lifetimes at higher dose rates at RT (*e.g.* Helliwell, 1988), this was only systematically investigated recently, when an inverse dose-rate effect was measured in-house between dose rates of 6 and 10 Gy s^{-1}, the higher rate giving four times the dose tolerance (*i.e.* four times the dose required to halve the total diffraction intensity, $D_{1/2}$) for chicken egg-white lysozyme crystals (Southworth-Davies *et al.*, 2007). For irradiation at a dose rate of 2800 Gy s^{-1} at a synchrotron at RT, ten times the dose tolerance has been recorded (Barker *et al.*, 2009).

Up until the 1990s, MX data were collected at RT, where the recommended practice was to monitor the intensity, I_0, of a particular reflection as the experiment proceeded, and to discard the crystal once the intensity had dropped to $0.85I_0$, or at the very worst, $0.70I_0$ if the particular crystals were in very short supply (Blundell & Johnson, 1976).

10.3.2. Cryocrystallography as a mitigation strategy

The search for strategies to reduce the rate of radiation damage in MX has resulted in the development over the last 20 years of cryocrystallographic techiques, whereby a crystal held by surface tension in a fibre, Mylar or polyimide loop is flash-cooled either into liquid nitrogen (77 K) or gaseous nitrogen at around 100 K (Garman, 1999; Garman & Schneider, 1997; see also Chapter 10.1). If the cooling is swift enough, the solvent around and within the crystal is present as amorphous solid, and does not form crystalline ice which would interfere with the diffraction and degrade the crystal order. To achieve this, cryoprotectant agents are added to the mother liquor, and optimum results can be obtained if the concentration of cryoprotectant results in equal shrinkage of the solvent channels and unit cell when the crystal is cooled (Juers & Matthews, 2001). At 100 K, nearly all the radical species in the crystal are immobilized and so, although the primary radiation damage is unchanged and the resulting electrons of higher energy, and perhaps holes, are still mobile, other radicals that are subsequently formed (*e.g.* hydroxyl and other oxygen-centred radicals) have significantly lower mobility and thus their potential to inflict damage is quenched.

Widespread use of these techniques seemed to overcome most radiation-damage problems, since cryocooling resulted in a vastly extended dose tolerance (around a factor of 70; Nave & Garman, 2005) for most macromolecular crystals. However, with the advent of more intense second- and third-generation synchrotron

Copyright © 2012 International Union of Crystallography

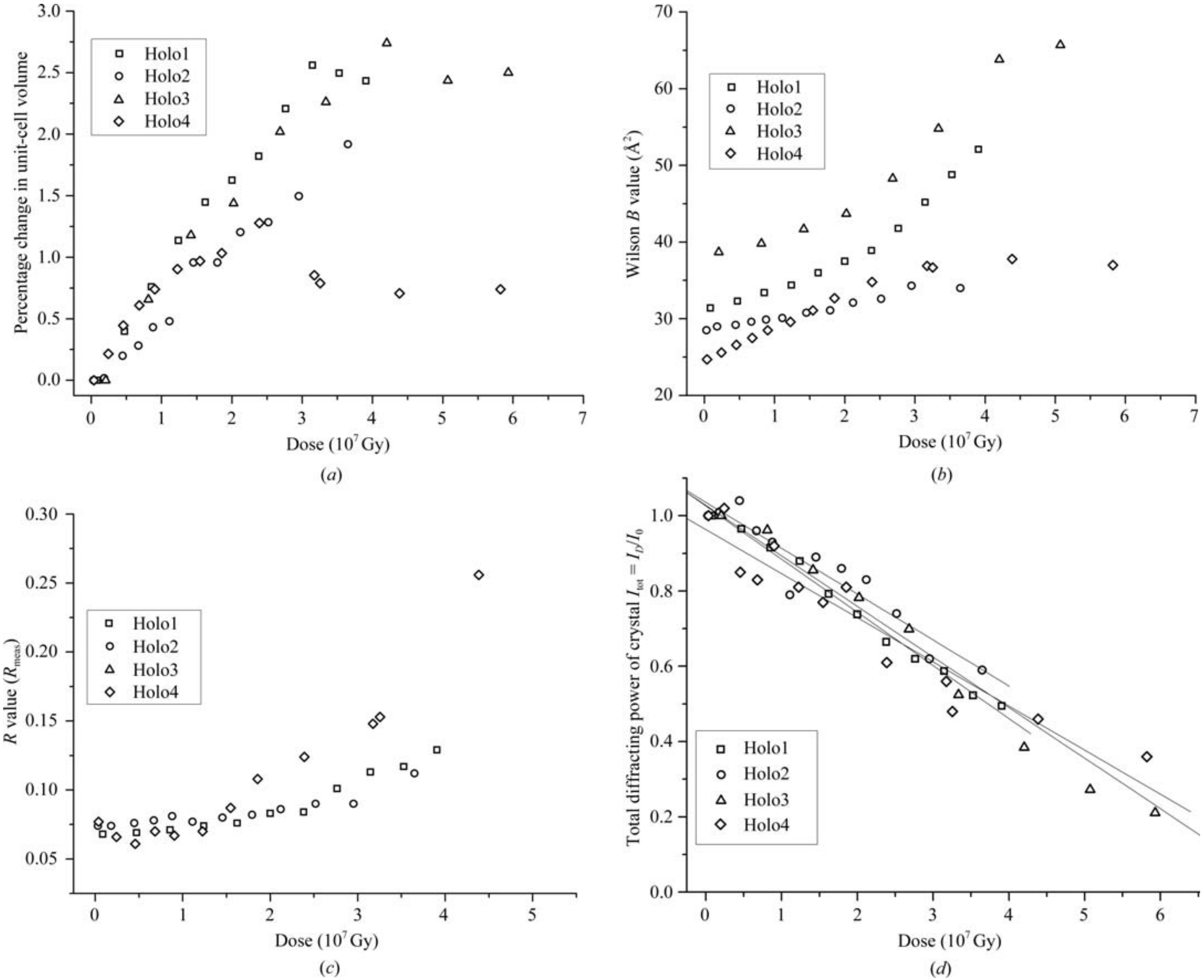


Figure 10.3.3.1

Global radiation-damage indicators as a function of dose for four holoferritin crystals (Owen *et al.*, 2006): (*a*) unit-cell volume, (*b*) Wilson *B* factor, (*c*) $R_{\rm merge}$ and (*d*) I(mean)/I_0(mean).

X-ray beams in the 1990s, observations of radiation damage to samples, even when held at 100 K, became common.

10.3.3. Characteristics of radiation damage at cryotemperatures

Systematic studies on this phenomenon have identified two separate indicators of damage as a function of dose: global (Fig. 10.3.3.1) and specific (Fig. 10.3.3.2) damage. The former results in a loss of the measured reflection intensities particularly at high resolution, expansion of the unit-cell volume, increasing values of the measure of the internal consistency of the data which quantifies the difference between reflection intensities that should ideally be the same ($R_{\rm merge}$), an increase in both the scaling *B* factors for the data and the atomic *B* factors of the refined structure, rotation of the molecule within the unit cell, and often (but not always) an increase in mosaicity. Visible changes in the samples are also observed (see Fig. 10.3.3.3). On warming following irradiation, bubbles of gas, thought to be hydrogen or CO_2, are emitted. Of more direct relevance to the biological interpretation of structures, specific structural damage to particular covalent bonds is observed to occur in a reproducible order in many proteins: first disulfide bridges are broken, then glutamates and aspartates are decarboxylated, tyrosine residues lose their hydroxyl group, and subsequently the carbon–sulfur bonds in methionines are cleaved (Burmeister, 2000; Ravelli & McSweeney, 2000; Weik *et al.*, 2000). Enzyme mechanisms can involve susceptible residues, so special care is required when interpreting structures that may have been modified by X-ray damage during the data collection. Metalloproteins are particularly vulnerable to partial reduction during the diffraction experiment and may not be in their native state by the end of the data collection (Carugo & Djinović Carugo, 2005). This specific damage often occurs well before there is any obvious degradation of the diffraction pattern.

In addition, as the diffraction experiment proceeds, the growing specific structural damage combined with the gradual increase in unit-cell volume and possible movement of the protein molecule within the unit cell induces creeping non-isomorphism on three simultaneous fronts. MAD (multiple-wavelength anomalous dispersion) structure solution thus becomes problematic, since by the time the third wavelength is collected, the cell and atomic structure can have changed such that the reflection intensities are significantly altered and this effect can obscure the anomalous signal required for structure solution.

The global effects at 100 K are thought to be dose-rate independent up to the flux densities currently used (10^{15} photons s^{-1} mm^{-2}) (Sliz *et al.*, 2003). Another study con-

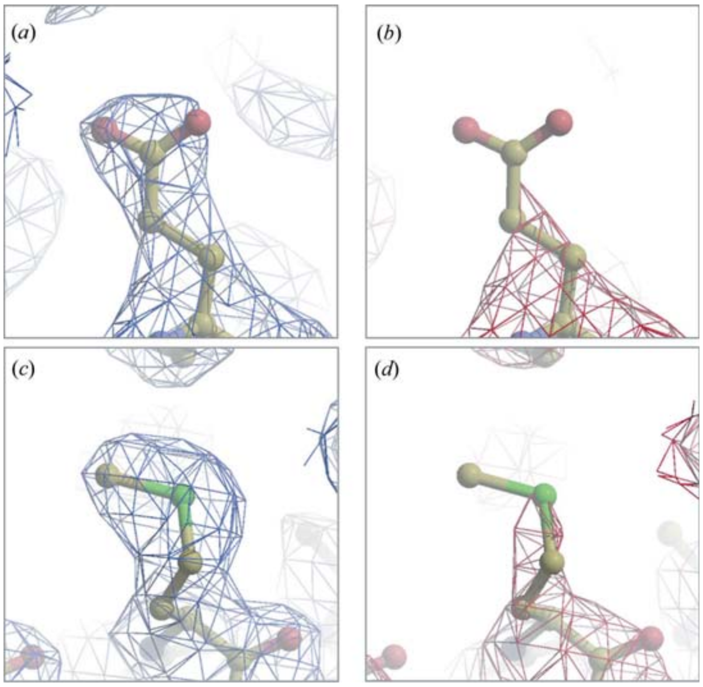


Figure 10.3.3.2
Specific structural damage inflicted on a cryocooled crystal of apoferritin during sequential data sets collected at ID14-4, ESRF: (*a*) $2F_o - F_c$ map of Glu63 contoured at 0.2 e Å^{-3} after a dose of 2.5 MGy and (*b*) after 50 MGy. (*c*) $2F_o - F_c$ map of Met96 contoured at 0.2 e Å^{-3} after a dose of 2.5 MGy and (*d*) after 50 MGy, showing loss of electron density around the disordered atoms (Garman & Owen, 2006).

curred with this finding, but indicated that there could be a second-order dose-rate effect based upon analysis of difference electron-density maps, specific damage being slightly more severe at higher dose rates (Leiros *et al.*, 2006). Conversely, however, Owen *et al.* (2006) reported a small (10%) reduction in $D_{1/2}$ for a dose-rate increase from 0.4×10^4 to 4.0×10^4 Gy s^{-1}.

Radiation damage thus ultimately results in lower-resolution structures, failed MAD structure solutions and sometimes the inaccurate interpretation of biological results if no control experiments are carried out to account for radiation-damage artefacts (Ravelli & Garman, 2006). It is thus an issue to be taken seriously by the structural biologist.

10.3.4. Understanding radiation damage

In order to monitor and understand radiation damage, it is vital to have reliable metrics to allow systematic and comparative studies. Several different global measures extracted from the data are currently being utilized: $[I/\sigma(I)]/[I_0/\sigma(I_0)]$, where I_0 is the initial intensity of a data set or of a set of reflections, $I(\text{mean})/I_0(\text{mean})$, the 'decay R factor', R_d, a pairwise R factor between identical and symmetry-related reflections occurring on different diffraction images, which is plotted against the difference in dose, ΔD, between the images in which the reflections were observed (Diederichs, 2006), isotropic B factor ($B_{\rm rel}$), which was found to be linearly dependent on dose, and a coefficient of sensitivity ($S_{\rm AD} = \Delta B_{\rm rel}/\Delta D 8\pi^2$, where $\Delta B_{\rm rel}/8\pi^2$ is the change in relative isotropic B factor and ΔD is the change in dose). The latter coefficient relates the increase in mean-squared atomic displacements to the dose, and was found to be a robust measure for radiation damage (Kmetko *et al.*, 2006). Unit-cell volume is not used, since it has been found to be an unreliable indicator (Murray & Garman, 2002; Ravelli *et al.*, 2002). Some of the above metrics are more appropriate than others for use in particular circumstances, *e.g.* if maximizing the resolution, trying to minimize radiation-damage-induced artefacts, or carrying out SAD (single-wavelength anomalous dispersion)/MAD structure solution. Care must be exercised in the calculation of electron-density maps extracted from sequential data sets, since the maps can be biased by the refinement, and so usually the phases from the refined structure obtained from the first data set are used for the map calculations of structures from the subsequent data sets.

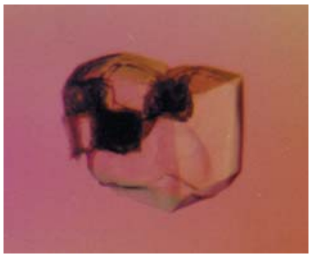


Figure 10.3.3.3
Photograph of a 400 µm neuraminidase crystal (subtype N9 from avian influenza isolated from a Noddy Tern) that has been irradiated on ID14-4 at the ESRF at 100 K and then allowed to warm to room temperature. The three black marks are from the 100 × 100 µm beam, the discolouration being an indication of radiation damage.

There is postulated to be a universal 'dose limit' for biological samples, beyond which they cannot be expected to survive intact and are destroyed by the energy deposited in them. From analogy with observations in electron microscopy (for 100 keV electrons), a dose limit of $D_{1/2} = 20$ MGy for a cryocooled (77 K) protein crystal was calculated by Henderson (1990). More recently Owen *et al.* (2006) have published an experimental study of the dose limit for cryocooled (100 K) protein crystals, in which it was found that the total intensity of diffraction had a linear dependence on absorbed dose (as opposed to the exponential dependence at room temperature). The dose required to halve the total intensity of the diffraction pattern was 43 MGy, but an upper limit of 30 MGy was suggested ($I = 0.7I_0$), beyond which point the biological information obtained from the experiment might be compromised. Coincidentally, this is the value of intensity loss previously recommended as the absolute maximum allowable for individual reflections during room-temperature data collection (Blundell & Johnson, 1976).

For systematic radiation-damage experiments and also for the 30 MGy experimental dose limit to be a useful predictive experimental tool, the dose absorbed by the crystal must be calculated. This can be conveniently computed for MX using the program *RADDOSE* (Murray *et al.*, 2004; Paithankar *et al.*, 2009), which requires information on the crystal, protein, cell and m*M* concentration of mother liquor, as well as the beam characteristics, and the duration and number of the exposures. The output gives the time available before the experimental dose limit is reached and the predicted temperature rise in the crystal using a simple isothermal 'lumped model'. Version 3 of *RADDOSE* takes into account the possible escape from the sample of fluorescence X-rays from the decay of excited atoms (only significant for atoms with atomic number greater than 20) following photoelectron emission, and also the energy deposited in the crystal by Compton scattering (non-negligible only above incident X-ray energies of 20 keV). If the crystal is bigger than the beam the program overestimates the dose, since the rotation of the crystal during the experiment is not taken into account and new un-irradiated material will be brought in with time, thus decreasing the dose.

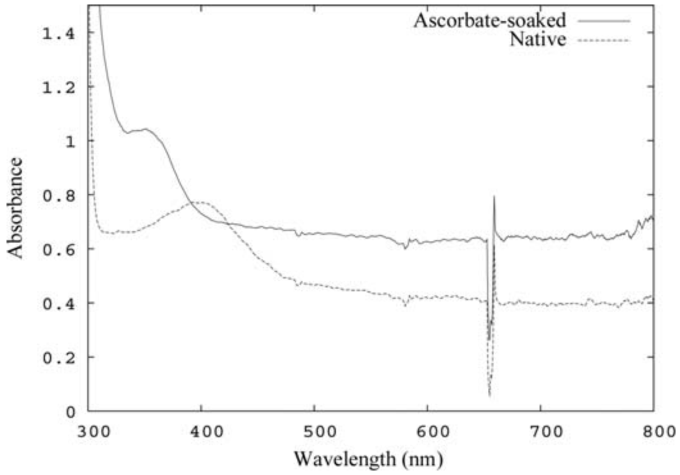


Figure 10.3.4.1
Microspectrophotometer absorption spectra of native and ascorbate-soaked hen egg-white lysozyme crystals at 100 K, showing the disulfide radical 400 nm peak formed on irradiation by a synchrotron X-ray beam, and the suppression of this peak on the addition of ascorbate, which has an absorption peak at 350 nm (Murray & Garman, 2002).

To account for the observed damage effects at 100 K, information must be gleaned from the radiation chemistry literature. Several analyses of the mobility of the various species formed upon X-ray exposure have been carried out and they help to explain the effectiveness of cryocooling for MX. Protons are only known to become mobile in amorphous ice above ~115 K (Fisher & Devlin, 1995) and although $^{\bullet}$OH radicals are trapped at 77 K in an aqueous glass of 6 *M* CsF (Becker *et al.*, 1994) and are inferred to be trapped in ice at 100 K but are thought to be mobile above 130 K and thus able to recombine (Symons, 1999), they have also been reported to be mobile at 77 K in a glass of DNA (Lange & Httermann, 1995). According to electron spin resonance measurements (Jones *et al.*, 1987), at 77 K positive holes in proteins are rapidly trapped, forming amido radicals on the protein backbone chain, whereas the electrons produced by inelastic interactions have significant mobility. Rao *et al.* (1983) showed that electrons added to proteins at 77 K are able to move efficiently until they encounter S—S bonds, where they are trapped.

Complementary methods are increasingly being employed for the observation of radiation damage in MX and specially designed instruments are being constructed on synchrotron beamlines to facilitate measurements performed simultaneously with X-ray diffraction, including UV–vis microspectroscopy (McGeehan *et al.*, 2009), UV–vis fluorescence, X-ray spectroscopy (Yano *et al.*, 2005) and Raman spectroscopy (Owen *et al.*, 2009). Optical peaks at 400 nm and between 550 and 600 nm are signatures of the formation of the disulfide radical anion, $RSSR^{\bullet}$ (Fig. 10.3.4.1) (Weik *et al.*, 2002), and the hydrated electron, respectively. The reduction of metal sites by the X-ray beam can also be observed online from their change in optical absorption (Hough *et al.*, 2008).

10.3.5. Mitigating and correcting for radiation damage

Various experimental parameters have been investigated to find ways of reducing the rate of radiation damage at cryotemperatures, including further reducing the temperature, changing the incident wavelength and adding radical scavengers.

Various studies have investigated the rate of global and specific damage infliction with temperature variation, and there is as yet no consensus on this issue. At 40 K, global indicators improved little (Teng & Moffat, 2002), but a study at 8 K found a 25% increase in dose tolerance (Chinte *et al.*, 2007). Another investigation, which also monitored global indicators, reported no improvement for insulin crystal dose tolerance at 15 K, but a 23% improvement for holoferritin, which has a large iron core (Meents *et al.*, 2007). However, significant protection against specific damage (a factor of 30 at 40 K compared to 110 K) has been observed as monitored by the intensity of the photoreduced peak in the X-ray absorption spectrum at the iron *K* edge on an iron-containing metalloprotein (Corbett *et al.*, 2007).

At current fluxes (4×10^{14} photons s^{-1} mm^{-2}), heating of the sample by the beam is predicted to be no more than 15 K (Mhaisekar *et al.*, 2005), so nitrogen cooling to 100 K is adequate to avoid the movement of species other than the already mobile electrons and perhaps positive holes (Jones *et al.*, 1987).

Despite much anecdotal evidence that longer-wavelength incident radiation causes swifter damage than shorter-wavelength incident radiation, systematic studies of this relationship (Shimizu *et al.*, 2007; Weiss *et al.*, 2005) have shown no detectable change in the damage rate with dose over the range 6.2 keV (2 Å) to 33 keV (0.275 Å). Although the absorption is higher at longer wavelengths, the probability of diffraction is also higher, and the ratio of scattering to dose is not predicted to change by more than 20% for a typical protein crystal (no heavy atoms) over the range of energies currently used in MX (7–20 keV).

Two scavengers have been tested at RT in the past (styrene and polyethylene glycol), but neither was found to be very effective and they have not been widely utilized. However, recently, 0.5 *M* ascorbate and 0.5 *M* 1,4-benzoquinone were reported to give an increase by factors of two and nine, respectively, in $D_{1/2}$ at RT. Most interestingly, for both scavengers, the dose dependence of the intensity decay was modified to a linear relationship rather than exhibiting the exponential intensity decay of the native crystal (Barker *et al.*, 2009). The decay of the total diffraction intensity observed at 100 K is predominantly linear (Fig. 10.3.3.1), and this result indicates that the radiation chemistry of the degradation of the crystalline order has been modified by the scavengers in a similar way as it is by temperature if cryocooled.

At 100 K, several scavengers have now been tested: styrene was reported to be ineffective but ascorbate reduced both global and specific damage (Murray & Garman, 2002). Nicotinic acid and 5,5′-dithiobis-2-nitrobenzoic acid (DTNB) have been found to be effective (Kauffmann *et al.*, 2006). A large number of potential scavengers were screened using an online microspectrophotometer (McGeehan *et al.*, 2009) for their ability to quench the formation of the disulfide radical anion 400 nm peak, and only ascorbate, 1,4-benzoquinone and TEMP (2,2,6,6,-tetramethyl-4-piperidone) eradicated it (Southworth-Davies & Garman, 2007). Thus, from the research carried out so far at 100 K, it seems that scavengers deserve further study but are not going to give a large improvement in dose tolerance and do not seem to be as potentially effective as they are at RT.

In addition to these experimental strategies, software development is underway to correct data for radiation damage ‘after the event’. These methods include zero-dose extrapolation, which uses multiple measurements of the same reflection to estimate its probable intensity at the beginning of the experiment (Blake & Phillips, 1962; Diederichs *et al.*, 2003), and the occupancy refinement of heavy-atom sites as a function of dose (Schiltz *et al.*, 2004).

10.3.6. Using radiation damage

The specific structural damage inflicted by radiation can be utilized for phasing, for example RIP, radiation-damage-induced phasing (Ravelli *et al.*, 2003), whereby a low-dose data collection (little specific damage) is followed by a 'burn', inflicted either by the X-ray beam or by a UV laser (Nanao & Ravelli, 2006), and then another low-dose data collection (disulfides broken and the relevant sulfur atoms thus disordered). This method has also been combined with anomalous scattering (RIPAS) (Zwart *et al.*, 2004).

Radiation damage can also be used to elucidate biological function and mechanism, and there are increasing numbers of such reports. For instance, the 'back-door' escape route of product clearance in acetylcholinesterase was identified using the X-ray beam to radiocleave a non-hydrolysable substrate analogue bound in the enzymatic site and collecting diffraction data sets at 100 and 150 K, the dynamic changes involved becoming visible at the higher temperature (Colletier *et al.*, 2008).

Other work exploiting radiation damage includes an inventive data-collection regime on a series of crystals utilizing the reduction of metal centres by the X-ray beam to track the resulting conformational changes, which shed light on the catalytic pathway of horseradish peroxidase (Berglund *et al.*, 2002).

10.3.7. Open questions

There are still many areas where systematic investigations are required to improve our understanding of the radiation chemistry within an irradiated protein crystal held at either room temperature or at various cryotemperatures, so that better strategies for minimizing damage can be developed. For instance, regarding specific damage, the varying susceptibility of the same type of residue within a protein is not yet understood: there appears to be no correlation with solvent accessibility (Fioravanti *et al.*, 2007). On the data-collection side, there is as yet no compelling evidence to warrant the routine use of helium cryostats (and the associated expense) apart from the case of metalloproteins, and there exists an urgent need to establish the best experimental strategies and optimum crystal sizes to allow the most advantageous use of newly available microbeams at synchrotrons.

The most useful contribution to be made by MX radiation-damage research is in identifying concrete experimental protocols for everyday use on synchrotron beamlines so that researchers can ensure that they obtain the maximum possible amount of high-quality data from their crystals. This would firstly facilitate structure solution, and secondly avoid compromising the biological information extracted from the structure once obtained.

References

Barker, A. I., Southworth-Davies, R. J., Paithankar, K. S., Carmichael, I. & Garman, E. F. (2009). *Room-temperature scavengers for macromolecular crystallography: increased lifetimes and modified dose dependence of the intensity decay*. *J. Synchrotron Rad.* **16**, 205–216.

Becker, D., La Vere, T. & Sevilla, M. D. (1994). *ESR detection at 77 K of the hydroxyl radical in the hydration layer of gamma-irradiated DNA*. *Radiat. Res.* **140**, 123–129.

Berglund, G. I., Carlsson, G. H., Smith, A. T., Szöke, H., Henriksen, A. & Hajdu, J. (2002). *The catalytic pathway of horseradish peroxidase at high resolution*. *Nature* (*London*), **417**, 463–468.

Blake, C. & Phillips, D. C. (1962). *Proceedings of the Symposium on the Biological Effects of Ionising Radiation at the Molecular Level*, pp. 183–191. Vienna: International Atomic Energy Agency.

Blundell, T. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Burmeister, W. P. (2000). *Structural changes in a cryo-cooled protein crystal owing to radiation damage*. *Acta Cryst.* D**56**, 328–341.

Carugo, O. & Djinović Carugo, K. (2005). *When X-rays modify the protein structure: radiation damage at work*. *Trends Biochem. Sci.* **30**, 213–219.

Chinte, U., Shah, B., Chen, Y.-S., Pinkerton, A. A., Schall, C. A. & Hanson, B. L. (2007). *Cryogenic (<20 K) helium cooling mitigates radiation damage to protein crystals*. *Acta Cryst.* D**63**, 486–492.

Colletier, J. P., Bourgeois, D., Sanson, B., Fournier, D., Sussman, J. L., Silman, I. & Weik, M. (2008). *Shoot-and-Trap: use of specific X-ray damage to study structural protein dynamics by temperature-controlled cryo-crystallography*. *Proc. Natl Acad. Sci. USA*, **105**, 11742–11747.

Corbett, M. C., Latimer, M. J., Poulos, T. L., Sevrioukova, I. F., Hodgson, K. O. & Hedman, B. (2007). *Photoreduction of the active site of the metalloprotein putidaredoxin by synchrotron radiation*. *Acta Cryst.* D**63**, 951–960.

Diederichs, K. (2006). *Some aspects of quantitative analysis and correction of radiation damage*. *Acta Cryst.* D**62**, 96–101.

Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Zero-dose extrapolation as part of macromolecular synchrotron data reduction*. *Acta Cryst.* D**59**, 903–909.

Fioravanti, E., Vellieux, F. M. D., Amara, P., Madern, D. & Weik, M. (2007). *Specific radiation damage to acidic residues and its relation to their chemical and structural environment*. *J. Synchrotron Rad.* **14**, 84–91.

Fisher, M. & Devlin, J. P. (1995). *Defect activity in amorphous ice from isotopic exchange data: insights into the glass transition*. *J. Phys. Chem.* **99**, 11584–11590.

Garman, E. (1999). *Cool data: quantity AND quality*. *Acta Cryst.* D**55**, 1641–1653.

Garman, E. F. & Owen, R. L. (2006). *Cryocooling and radiation damage in macromolecular crystallography*. *Acta Cryst.* D**62**, 32–47.

Garman, E. F. & Schneider, T. R. (1997). *Macromolecular cryocrystallography*. *J. Appl. Cryst.* **30**, 211–237.

Helliwell, J. R. (1988). *Protein crystal perfection and the nature of radiation damage*. *J. Cryst. Growth*, **90**, 259–272.

Henderson, R. (1990). *Cryo-protection of protein crystals against radiation damage in electron and X-ray diffraction*. *Proc. R. Soc. London Ser. B*, **241**, 6–8.

Hough, M. A., Antonyuk, S. V., Strange, R. W., Eady, R. R. & Hasnain, S. S. (2008). *Crystallography with online optical and X-ray absorption spectroscopies demonstrates an ordered mechanism in copper nitrite reductase*. *J. Mol. Biol.* **378**, 353–361.

Jones, G. D., Lea, J. S., Symons, M. C. & Taiwo, F. A. (1987). *Structure and mobility of electron gain and loss centres in proteins*. *Nature* (*London*), **330**, 772–773.

Juers, D. H. & Matthews, B. W. (2001). *Reversible lattice repacking illustrates the temperature dependence of macromolecular interactions*. *J. Mol. Biol.* **311**, 851–862.

Kauffmann, B., Weiss, M. S., Lamzin, V. S. & Schmidt, A. (2006). *How to avoid premature decay of your macromolecular crystal: a quick soak for long life*. *Structure*, **14**, 1099–1105.

Kmetko, J., Husseini, N. S., Naides, M., Kalinin, Y. & Thorne, R. E. (2006). *Quantifying X-ray radiation damage in protein crystals at cryogenic temperatures*. *Acta Cryst.* D**62**, 1030–1038.

Lange, M. B. W. & Httermann, J. (1995). *Influence of electron scavengers on the radical formation in thymidine-5′-monophosphate and DNA in frozen aqueous solution and glasses*. *Int. J. Radiat. Biol.* **68**, 475–486.

Leiros, H.-K. S., Timmins, J., Ravelli, R. B. G. & McSweeney, S. M. (2006). *Is radiation damage dependent on the dose rate used during macromolecular crystallography data collection?* *Acta Cryst.* D**62**, 125–132.

McGeehan, J., Ravelli, R. B. G., Murray, J. W., Owen, R. L., Cipriani, F., McSweeney, S., Weik, M. & Garman, E. F. (2009). *Colouring cryo-cooled crystals: online microspectrophotometry*. *J. Synchrotron Rad.* **16**, 163–172.

Meents, A., Wagner, A., Schneider, R., Pradervand, C., Pohl, E. & Schulze-Briese, C. (2007). *Reduction of X-ray-induced radiation damage of macromolecular crystals by data collection at 15 K: a systematic study*. *Acta Cryst.* D**63**, 302–309.

Mhaisekar, A., Kazmierczak, M. J. & Banerjee, R. (2005). *Three-dimensional numerical analysis of convection and conduction cooling of spherical biocrystals with localized heating from synchrotron X-ray beams*. *J. Synchrotron Rad.* **12**, 318–328.

Murray, J. & Garman, E. (2002). *Investigation of possible free-radical scavengers and metrics for radiation damage in protein cryocrystallography*. *J. Synchrotron Rad.* **9**, 347–354.

Murray, J. W., Garman, E. F. & Ravelli, R. B. G. (2004). *X-ray absorption by macromolecular crystals: the effects of wavelength and crystal composition on absorbed dose*. *J. Appl. Cryst.* **37**, 513–522.

Nanao, M. H. & Ravelli, R. B. (2006). *Phasing macromolecular structures with UV-induced structural changes*. *Structure*, **14**, 791–800.

Nave, C. & Garman, E. F. (2005). *Towards an understanding of radiation damage in cryocooled macromolecular crystals*. *J. Synchrotron Rad.* **12**, 257–260.

O'Neill, P., Stevens, D. L. & Garman, E. (2002). *Physical and chemical considerations of damage induced in protein crystals by synchrotron radiation: a radiation chemical perspective*. *J. Synchrotron Rad.* **9**, 329–332.

Owen, R. L., Pearson, A. R., Meents, A., Boehler, P., Thominet, V. & Schulze-Briese, C. (2009). *A new on-axis multimode spectrometer for the macromolecular crystallography beamlines of the Swiss Light Source*. *J. Synchrotron Rad.* **16**, 173–182.

Owen, R. L., Rudino-Pinera, E. & Garman, E. F. (2006). *Experimental determination of the radiation dose limit for cryocooled proteins*. *Proc. Natl Acad. Sci. USA*, **103**, 4912–4917.

Paithankar, K. S., Owen, R. L. & Garman, E. F. (2009). *Absorbed dose calculations for macromolecular crystals: improvements to RADDOSE*. *J. Synchrotron Rad.* **16**, 152–162.

Rao, D. N., Symons, M. C. R. & Stephenson, J. M. (1983). *Radiation-induced electron capture by proteins containing disulfide linkages – an electron-spin resonance study*. *J. Chem. Soc. Perkin Trans. 2*, pp. 727–730.

Ravelli, R. B., Leiros, H. K., Pan, B., Caffrey, M. & McSweeney, S. (2003). *Specific radiation damage can be used to solve macromolecular crystal structures*. *Struct. Fold Des.* **11**, 217–224.

Ravelli, R. B. G. & Garman, E. (2006). *Radiation damage in macromolecular crystallography*. *Curr. Opin. Struct. Biol.* **16**, 624–629.

Ravelli, R. B. G. & McSweeney, S. M. (2000). *The 'fingerprint' that X-rays can leave on structures*. *Struct. Fold Des.* **8**, 315–328.

Ravelli, R. B. G., Theveneau, P., McSweeney, S. & Caffrey, M. (2002). *Unit-cell volume change as a metric of radiation damage in crystals of macromolecules*. *J. Synchrotron Rad.* **9**, 355–360.

Schiltz, M., Dumas, P., Ennifar, E., Flensburg, C., Paciorek, W., Vonrhein, C. & Bricogne, G. (2004). *Phasing in the presence of severe site-specific radiation damage through dose-dependent modelling of heavy atoms*. *Acta Cryst.* D**60**, 1024–1031.

Shimizu, N., Hirata, K., Hasegawa, K., Ueno, G. & Yamamoto, M. (2007). *Dose dependence of radiation damage for protein crystals studied at various X-ray energies*. *J. Synchrotron Rad.* **14**, 4–10.

Sliz, P., Harrison, S. C. & Rosenbaum, G. (2003). *How does radiation damage in protein crystals depend on X-ray dose? Structure*, **11**, 13–19.

Southworth-Davies, R. J. & Garman, E. F. (2007). *Radioprotectant screening for cryocrystallography*. *J. Synchrotron Rad.* **14**, 73–83.

Southworth-Davies, R. J., Medina, M. A., Garman, E. F. & Carmichael, I. (2007). *Observation of decreased radiation damage at higher dose rates in room temperature protein crystallography*. *Structure*, **15**, 1531–1541.

Symons, M. C. R. (1999). *Mechanism of radiation damage to proteins and DNA – an EPR perspective*. *Prog. React. Kinet. Mech.* **24**, 139164.

Teng, T.-Y. & Moffat, K. (2002). *Radiation damage of protein crystals at cryogenic temperatures between 40 K and 150 K*. *J. Synchrotron Rad.* **9**, 198–201.

Weik, M., Bergès, J., Raves, M. L., Gros, P., McSweeney, S., Silman, I., Sussman, J. L., Houée-Levin, C. & Ravelli, R. B. G. (2002). *Evidence for the formation of disulfide radicals in protein crystals upon X-ray irradiation*. *J. Synchrotron Rad.* **9**, 342–346.

Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Specific chemical and structural damage to proteins produced by synchrotron radiation*. *Proc. Natl Acad. Sci. USA*, **97**, 623–628.

Weiss, M. S., Panjikar, S., Mueller-Dieckmann, C. & Tucker, P. A. (2005). *On the influence of the incident photon energy on the radiation damage in crystalline biological samples*. *J. Synchrotron Rad.* **12**, 304–309.

Yano, J., Kern, J., Irrgang, K. D., Latimer, M. J., Bergmann, U., Glatzel, P., Pushkar, Y., Biesiadka, J., Loll, B., Sauer, K., Messinger, J., Zouni, A. & Yachandra, V. K. (2005). *X-ray damage to the Mn4Ca complex in single crystals of photosystem II: a case study for metalloprotein crystallography*. *Proc. Natl Acad. Sci. USA*, **102**, 12047–12052.

Zwart, P. H., Banumathi, S., Dauter, M. & Dauter, Z. (2004). *Radiation-damage-induced phasing with anomalous scattering: substructure solution and phasing*. *Acta Cryst.* D**60**, 1958–1963.

International Tables for Crystallography (2012). Vol. F, Chapter 11.1, pp. 263–265.

PART 11. DATA PROCESSING

Chapter 11.1. Automatic indexing of oscillation images

M. G. Rossmann

11.1.1. Introduction

Auto-indexing routines have been used extensively for initiating diffraction data collection with a single-point-detector device (Sparks, 1976, 1982). These methods depend upon the precise knowledge of the reciprocal-lattice vectors for a few selected reflections. Greater difficulty has been encountered for automatic indexing of oscillation images recorded on two-dimensional detectors using randomly oriented crystals, as is frequently the case for macromolecular crystal samples. In the past, the practice was to orient crystals relative to the camera axes with an accuracy of at least 1°. In this case, the indexing procedure required only refinement of the crystal orientation matrix (Wonacott, 1977; Rossmann, 1979). The 'American method' (Rossmann & Erickson, 1983), where crystals are oriented more or less randomly, is currently used because of the need for optimizing available synchrotron time and because of the deterioration in radiation-sensitive crystals during the setting process.

A variety of techniques were suggested to determine the crystal orientation, some of which required initial knowledge of the cell dimensions (Vriend & Rossmann, 1987; Kabsch, 1988), while more advanced techniques (Kim, 1989; Higashi, 1990; Kabsch, 1993) determined both cell dimensions and crystal orientation. All these methods start with the determination of the reciprocal-lattice vectors assuming that the oscillation photographs are 'stills'. The methods of Higashi and Kabsch, as well as, in part, Kim's, analyse the distribution of the difference vectors generated from the reciprocal-lattice vectors. The most frequent difference vectors are taken as the basis vectors defining the reciprocal-lattice unit cell and its orientation. In addition, Kim's technique requires the input of the orientation of a likely zone-axis direction onto which the reciprocal-lattice vectors are then projected. The projections will have a periodicity distribution consistent with the reciprocal-lattice planes perpendicular to the zone axis. Duisenberg (1992) used a similar approach for single-point-detector data, although he did not rely on prior knowledge of the zone-axis direction. Instead, he defined possible zone axes as being perpendicular to a reciprocal-lattice plane by combining three, suitably chosen, reciprocal-lattice points.

None of the above techniques were entirely satisfactory as they sometimes failed to find a suitable crystal orientation matrix. A major advance was made in the program *DENZO*, a part of the *HKL* package (Otwinowski & Minor, 1997), which not only has a robust indexing procedure but also has a useful graphical interface. Unfortunately, the indexing technique used in the procedure has never been described, except for a few hints in the manual on the use of an FFT (fast Fourier transform). Indeed, Bricogne (1986) suggested that a three-dimensional Fourier transformation might be a powerful indexing tool, and Strouse (1996) developed such a procedure for single-point-detector data. However, for large unit cells this procedure requires an excessive amount of memory and time (Campbell, 1997).

11.1.2. The crystal orientation matrix

The position $\mathbf{x}\ (x, y, z)$ of a reciprocal-lattice point can be given as

$$\mathbf{x} = [\Phi][A]\mathbf{h}. \qquad (11.1.2.1)$$

The matrix $[\Phi]$ is a rotation matrix around the camera's spindle axis for a rotation of φ. The vector $\mathbf{h}$ represents the Miller indices (h, k, l) and $[A]$ defines the reciprocal unit-cell dimensions and the orientation of the crystal lattice with respect to the camera axes when $\varphi = 0$. Thus,

$$[A] = \begin{pmatrix} a_x^* & b_x^* & c_x^* \\ a_y^* & b_y^* & c_y^* \\ a_z^* & b_z^* & c_z^* \end{pmatrix}, \qquad (11.1.2.2)$$

where a_x^*, a_y^* and a_z^* are the components of the crystal $\mathbf{a}^*$ axis with respect to the orthogonal camera axes. When an oscillation image is recorded, the position of a reciprocal-lattice point is moved from $\mathbf{x}_1$ to $\mathbf{x}_2$, corresponding to a rotation of the crystal from φ_1 to φ_2. The recorded position of the reflection on the detector corresponds to the point $\mathbf{x}$ when it is on the Ewald sphere somewhere between $\mathbf{x}_1$ and $\mathbf{x}_2$. The actual value of φ at which this crossing occurs cannot be retrieved from the oscillation image. We shall therefore assume here, as is the case in all other procedures, that $[\Phi][A]$ defines the crystal orientation in the centre of the oscillation range. Defining the camera axes as in Rossmann (1979), it is easy to show that a reflection recorded at the position (X, Y) on a flat detector normal to the X-ray beam, at a distance D from the crystal, corresponds to

$$\begin{aligned} x &= \frac{X}{\lambda(X^2 + Y^2 + D^2)^{1/2}} \\ y &= \frac{Y}{\lambda(X^2 + Y^2 + D^2)^{1/2}} \\ z &= \frac{D}{\lambda(X^2 + Y^2 + D^2)^{1/2}}, \end{aligned} \qquad (11.1.2.3)$$

where λ is the X-ray wavelength.

If an approximate $[A]$ matrix is available, the Miller indices of an observed peak at (X, Y) can be roughly determined using (11.1.2.3) and (11.1.2.1), where

$$\mathbf{h} = [A]^{-1}[\Phi]^{-1}\mathbf{x}, \qquad (11.1.2.4)$$

with the error being dependent upon the width of the oscillation range, the error in the detector parameters and errors in determining the coordinates of the centres of the recorded reflections.

11.1.3. Fourier analysis of the reciprocal-lattice vector distribution when projected onto a chosen direction

If the members of a set of reciprocal-lattice planes perpendicular to a chosen direction are well separated, then the projections of the reciprocal-lattice vectors onto this direction will have an easily recognizable periodic distribution (Fig. 11.1.3.1). Unlike the procedure of Kim (1989), which requires the input of a likely

Copyright © 2012 International Union of Crystallography

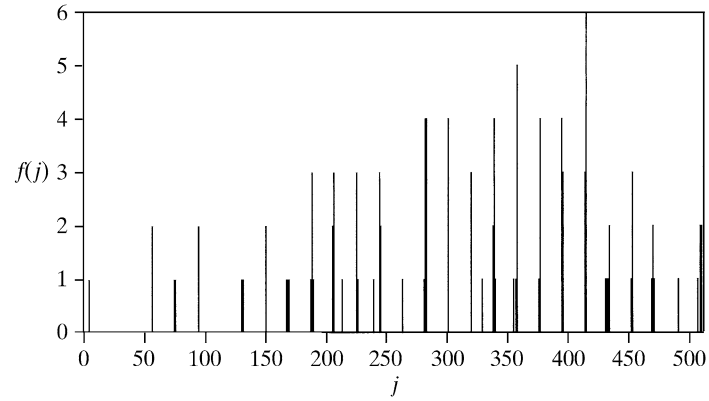


Figure 11.1.3.1
Frequency distribution of the projected reciprocal-lattice vectors for a suitably chosen direction of a diffraction pattern from a fibritin crystal (Tao *et al.*, 1997). Reproduced with permission from Steller *et al.* (1997).

zone-axis direction, the present procedure tests all possible directions and analyses the frequency distribution of the projected reciprocal-lattice vectors in each case. Also, unlike the procedure of Kim, the periodicity is determined using an FFT.

Let **t** represent a dimensionless unit vector of a chosen direction. Then, the projection p of a reciprocal-lattice point **x** onto the chosen vector **t** is given by

$$p = \mathbf{x} \cdot \mathbf{t}. \qquad (11.1.3.1)$$

To apply a discrete FFT algorithm, all such projections of the reciprocal-lattice points onto the chosen direction **t** are sampled in small increments of p. For the given direction, the values of the projections are in a range between the endpoints $p_{\min}$ and $p_{\max}$. If the maximum real cell dimension is assumed to be $a_{\max}$, then the maximum number of reciprocal-lattice planes between the observed limits of p is $(p_{\max} - p_{\min})/(1/a_{\max})$. Hence, the number of useful grid points along the direction **t** should be

$$m = (p_{\max} - p_{\min}) n a_{\max}, \qquad (11.1.3.2)$$

where n represents the number of grid points between successive reciprocal-lattice planes and is normally set to 5. Then, the frequency $f(p)$ in the range $p < \mathbf{x} \cdot \mathbf{t} < p + \Delta p$ can be given as $f(p)\Delta p = f(j)$, where j is the closest integer to $(p - p_{\min})/\Delta p$ and $\Delta p = n a_{\max}$. Thus, the discrete Fourier transform of this frequency distribution will be given by the summation

$$F(k) = \sum_{j=0}^{m} f(j) \exp(2\pi i k j). \qquad (11.1.3.3)$$

The transform is then calculated using a fast Fourier algorithm for all integer values between 0 and $m/2$ (Fig. 11.1.3.2). The Fourier coefficients that best represent the periodicity of the frequency distribution will be large. The largest coefficient will occur at $k = 0$ and correspond to the number of vectors used in establishing the frequency distribution. The next set of large coefficients will correspond to the periodicity that represents every reciprocal-lattice plane. The ratio of this maximum to $F(0)$ will be a measure of the tightness of the frequency distribution around each lattice plane. Subsequent maxima will be due to periodicities spanning every second, third *etc.* frequency maximum and will thus be progressively smaller (Fig. 11.1.3.2). The largest $F(k)$ (when $k = l$), other than $F(0)$, will, therefore, correspond to an interval of d^* between reciprocal-lattice planes in the direction of **t** where $d^* = l/(na_{\max})$.

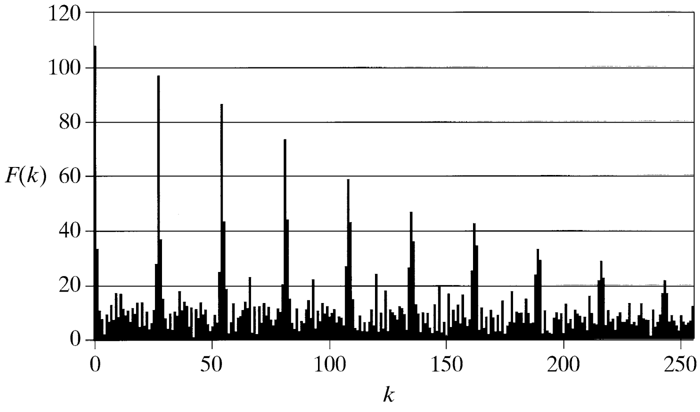


Figure 11.1.3.2
Fourier analysis of the distribution shown in Fig. 11.1.3.1. The first maximum, other than $F(0)$, is at $k = 27$, corresponding to $(1/d^*) = 41.9$ Å and a value of $F(27) = 97.0$. Reproduced with permission from Steller *et al.* (1997).

11.1.4. Exploring all possible directions to find a good set of basis vectors

The polar coordinates ψ, φ will be used to define the direction **t**, where ψ defines the angle between the X-ray beam and the chosen direction **t**. The Fourier analysis is performed for each direction **t** in the range $0 < \psi \leq \pi/2, 0 < \varphi \leq 2\pi$. A suitable angular increment in ψ was determined empirically to be about 0.03 rad (1.7°). For each value of ψ, the increment in φ is taken to be the closest integral value to $(2\pi \sin \psi)/0.03$. This procedure results in ~7300 separate, roughly equally spaced, directions.

For each direction **t**, the distribution of the corresponding $F(k)$ coefficients is surveyed to locate the largest local maximum at $k = l$. The ψ and φ values associated with the 30 largest maxima are selected for refinement by a local search procedure to obtain an accuracy of 10^{-4} rad (~0.006°). If the initial angular increment (0.03 rad) used for the hemisphere search was reduced, then it would not be necessary to refine quite as many local maxima. However, to increase the efficiency of the search procedure, the ratio of angular increments to the number of refined positions was chosen to minimize the total computing time. The $F(l)$ values of the refined positions are then sorted by size. Directions are chosen from these vectors to give a linearly independent set of three basis vectors of a primitive real-space unit cell. These are then converted to the basis vectors of the reciprocal cell. The components of the three reciprocal-cell axes along the three camera axes are the nine components of the crystal orientation matrix $[A]$ (11.1.2.2).

The final step in the selection of the best $[A]$ matrix is to choose various nonlinear combinations of the refined vectors that have the biggest $F(l)$ values. That set of three vectors which gives the best indexing results is then chosen to represent the crystal orientation matrix $[A]$. A useful criterion is to determine the nonintegral Miller indices h' from (11.1.2.4) using the $[A]$ matrix and the known reciprocal-lattice vectors **x**. Any reflection for which any component $|\mathbf{h} - \mathbf{h}'|$ is bigger than, say, 0.2 is rejected. The best $[A]$ matrix is chosen as the one with the least number of rejections. In most cases, the best combination corresponds to taking the three largest $F(l)$ values.

The program goes on to determine a reduced cell from the cell obtained by the above indexing procedure (Kim, 1989). The reduced cell is then analysed in terms of the 44 lattice characters (Burzlaff *et al.*, 2005; Kabsch, 1993) in order to evaluate the most likely Bravais lattice and crystal system.

11.1.5. The program

The auto-indexing program has been written in C and implemented on an SGI O^2 workstation. It is a component of the general data-processing system (*DPS*). The auto-indexing component of *DPS* is available over the web, including the source code. The run-time for auto-indexing is sufficiently short for the procedure to be run interactively. The program described here is now included in the *MOSFLM* data-processing package (Leslie, 1992).

This short article is adapted from a fuller description of the auto-indexing procedure (Steller *et al.*, 1997). I thank Cheryl Towell and Sharon Wilder for help in the preparation of this manuscript. The work was supported by a grant from the National Science Foundation to MGR. Ingo Steller, who wrote and developed the program, was supported by a Deutsche Forschungsgemeinschaft postdoctoral fellowship.

References

Bricogne, G. (1986). *Indexing and the Fourier transform.* In *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase III)*, p. 28. LURE, Paris, 12–19 November.

Burzlaff, H., Gruber, B., Zimmermann, H. & de Wolff, P. M. (2005). *Crystal lattices*. In *International Tables for Crystallography*, Vol. A. *Space-Group Symmetry*, edited by Th. Hahn, Part 9. Heidelberg: Springer.

Campbell, J. W. (1997). In *CCP4 Newsletter*, No. 33, edited by M. Winn. Warrington: Daresbury Laboratory.

Duisenberg, A. J. M. (1992). *Indexing in single-crystal diffractometry with an obstinate list of reflections. J. Appl. Cryst.* **25**, 92–96.

Higashi, T. (1990). *Auto-indexing of oscillation images. J. Appl. Cryst.* **23**, 253–257.

Kabsch, W. (1988). *Automatic indexing of rotation diffraction patterns. J. Appl. Cryst.* **21**, 67–71.

Kabsch, W. (1993). *Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. J. Appl. Cryst.* **26**, 795–800.

Kim, S. (1989). *Auto-indexing oscillation photographs. J. Appl. Cryst.* **22**, 53–60.

Leslie, A. G. W. (1992). *Recent changes to the MOSFLM package for processing film and image plate data.* In *CCP4 and ESF-EACMB Newsletter on Protein Crystallography*, No. 26. Warrington: Daresbury Laboratory.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol.* **276**, 307–326.

Rossmann, M. G. (1979). *Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. J. Appl. Cryst.* **12**, 225–238.

Rossmann, M. G. & Erickson, J. W. (1983). *Oscillation photography of radiation-sensitive crystals using a synchrotron source. J. Appl. Cryst.* **16**, 629–636.

Sparks, R. A. (1976). *Trends in minicomputer hardware and software. Part I.* In *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedláček, pp. 452–467. Copenhagen: Munksgaard.

Sparks, R. A. (1982). *Data collection with diffractometers.* In *Computational Crystallography*, edited by D. Sayre, pp. 1–18. New York: Oxford University Press.

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *An algorithm for automatic indexing of oscillation images using Fourier analysis. J. Appl. Cryst.* **30**, 1036–1040.

Strouse, C. E. (1996). Personal communication.

Tao, Y., Strelkov, S. V., Mesyanzhinov, V. V. & Rossmann, M. G. (1997). *Structure of bacteriophage T4 fibritin: a segmented coiled coil and the role of the C-terminal domain. Structure*, **5**, 789–798.

Vriend, G. & Rossmann, M. G. (1987). *Determination of the orientation of a randomly placed crystal from a single oscillation photograph. J. Appl. Cryst.* **20**, 338–343.

Wonacott, A. J. (1977). *Geometry of the rotation method.* In *The Rotation Method in Crystallography*, edited by U. W. Arndt & A. J. Wonacott, pp. 75–103. Amsterdam: North-Holland Publishing Co.

references

International Tables for Crystallography (2012). Vol. F, Chapter 11.2, pp. 266–271.

Chapter 11.2. Integration of macromolecular diffraction data

A. G. W. Leslie

11.2.1. Introduction

Data integration refers to the process of obtaining estimates of diffracted intensities (and their standard deviations) from the raw images recorded by an X-ray detector. As two-dimensional (2D) area detectors are almost universally used to collect macromolecular diffraction data, only this type of detector will be considered in the following analysis.

When collecting data with a 2D area detector, a decision has to be taken about the magnitude of the angular rotation of the crystal during the recording of each image. Two distinct modes of operation are possible: the rotation per image can be comparable to, or greater than, the angular reflection range of a typical reflection (coarse φ slicing), or it can be much less than the reflection width (fine φ slicing). The latter approach allows the use of three-dimensional profile fitting and, providing that the detector is relatively noise-free, improves the quality of the resulting data by minimizing the contribution of the X-ray background to the total measured intensity. However, there are significant overheads associated with recording, storing and processing the relatively large number of images that are required. Three-dimensional profile fitting is described in Chapter 11.3 and will not be discussed here.

11.2.2. Prerequisites for accurate integration

11.2.2.1. Crystal parameters

Only the integration procedure itself will be described in detail in this article. However, in order to obtain the highest quality data possible from a given set of images, there are a number of parameters that need to be determined in advance of, or during, the integration. The most important of these are the unit-cell parameters , which should be determined to an accuracy of a few parts in a thousand (or better). Post-refinement procedures (Winkler *et al.*, 1979; Rossmann *et al.*, 1979), which make use of the estimated φ centroids of observed spots rather than their detector coordinates, generally provide more accurate estimates than methods based on the spot positions. This is because spot positions are affected by residual spatial distortions (after applying appropriate corrections) and the cell parameters are correlated with the crystal-to-detector distance, which is not always accurately known. For either method, it is necessary to include data from widely separated regions of reciprocal space (ideally φ values 90° apart) in order to determine all unit-cell parameters accurately. This is particularly important for lower-symmetry space groups.

The crystal orientation also needs to be known to an accuracy that corresponds to a few per cent of the reflection width. For crystals with low mosaicity (*e.g.* 0.1°) this corresponds to a hundredth of a degree or better. Fortunately, it is a feature of post refinement that the error in determining the orientation is typically a few per cent of the reflection width, and so this condition can generally be met. It is important to allow for movement of the crystal by continuously updating the crystal orientation during integration. This is even true when using cryo-cooled crystals, as the magnetic couplings that attach the pin (holding the crystal) to the goniometer head are not strong enough to prevent small movements, particularly with the high angular rotation rates employed on intense synchrotron beamlines. Non-orthogonality of the incident X-ray beam and the rotation axis (if not allowed for) or an off-centre crystal will also give rise to apparent changes in crystal orientation with spindle rotation.

The crystal mosaicity can be estimated by visual inspection and refined by post refinement. Refined values are quite reliable when the mosaic spread is less than about 0.5°, but become more dependent on the rocking-curve model for the high mosaicities that are often associated with frozen crystals. The presence of diffuse scatter, which appears as haloes around the Bragg diffraction spots, presents further difficulties in determining the correct mosaic spread. When processing coarse-sliced images it is preferable to overestimate the mosaic spread slightly (rather than underestimate it). This will result in an increase in random errors (by adding in the X-ray background from an image on which the spot is not actually present), whereas using too small a value can give systematic errors (by underestimating the number of images on which the spot lies).

11.2.2.2. Detector parameters

Detector calibration is essential for high data quality. Both the spatial distortion and the non-uniformity of response of the detector must be accurately known, and it is equally important that these corrections are stable over the timescale of the experiment (and preferably for much longer).

Finally, the crystal-to-detector distance, the detector orientation and the direct-beam position must be refined and continuously updated during integration, using observed spot positions. The crystal-to-detector distance can vary during data collection if the crystal is not exactly centred on the rotation axis, and the direct-beam position can move after a beam refill at a synchrotron. For image-plate detectors with two (or more) plates, the direct-beam position and detector distance often differ slightly for different plates.

With appropriate care, it is normally possible to predict reflection positions on the detector to an accuracy of 20–30 µm, or a fraction of the pixel size, particularly for highly collimated X-ray beams available at synchrotron sources. This level of accuracy is necessary to minimize possible systematic errors, particularly in the case of profile fitting.

11.2.3. Methods of integration

There are two quite distinct procedures available for determining the integrated intensities: summation integration and profile fitting. Summation integration involves simply adding the pixel values for all pixels lying within the area of a spot, and then subtracting the estimated background contribution to the same pixels. Profile fitting (Diamond, 1969; Ford, 1974; Rossmann, 1979) assumes that the actual spot shape or profile is known (in two or three dimensions) and the intensity is derived by finding the scale factor that, when applied to the known (or standard) profile, gives the best fit to the observed spot profile. In practice,

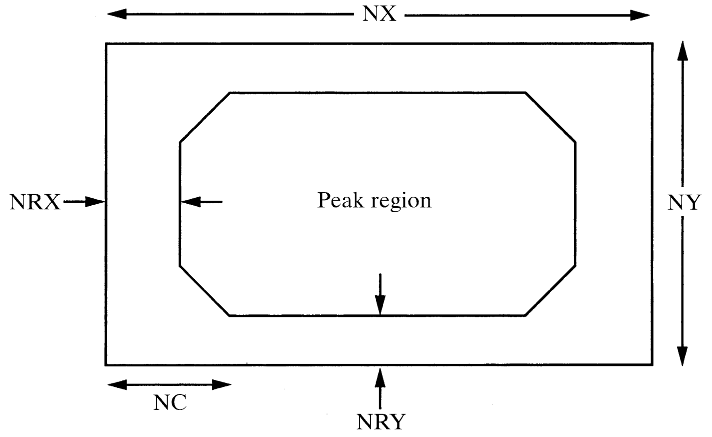


Figure 11.2.4.1
The measurement-box definition used in *MOSFLM*. The measurement box has overall dimensions of NX by NY pixels (both odd integers). The separation between peak and background pixels is defined by the widths of the background rims (NRX and NRY) and the corner cutoff (NC). The size of the peak region is optimized separately for each of the standard profiles.

profile fitting requires two separate steps: the determination of the standard profiles and the evaluation of the profile-fitting intensities. As will be shown later, profile fitting results in a reduction in the random error associated with weak intensities, but offers no improvement for very high intensities.

11.2.4. The measurement box

X-ray scattering from air, the sample holder and the specimen itself gives rise to a general background in the images which has to be subtracted in order to obtain the Bragg intensities. Ideally, the background should be measured for the same pixels used to record the Bragg diffraction spot, but this is not usually practical and the background is determined using pixels immediately adjacent to the spot. In practice, the pixels to be used for the determination of the background (background pixels) and those to be used for evaluating the intensity (peak pixels) are defined using a 'measurement box'. This is a rectangular box of pixels centred on the predicted spot position. Each pixel within the box is classified as being a background or a peak pixel (or neither). This mask can either be defined by the user, or the classification can be made automatically by the program. An example of a possible measurement-box definition is given in Fig. 11.2.4.1. The background parameters NRX, NRY and NC can be optimized automatically by maximizing the ratio of the intensity divided by its standard deviation, in a manner analogous to that described by Lehmann & Larsen (1974). It is generally assumed that the background can be adequately modelled as a plane, and the plane constants are determined using the background pixels. This allows the background to be estimated for the peak pixels, so that the background-corrected intensity can be calculated.

11.2.5. Integration by simple summation

11.2.5.1. Determination of the best background plane

The background plane constants a, b, c are determined by minimizing

$$R_1 = \sum_{i=1}^{n} w_i(\rho_i - ap_i - bq_i - c)^2, \qquad (11.2.5.1)$$

where ρ_i is the total counts at the pixel with coordinates $(p_i q_i)$ with respect to the centre of the measurement box, and the summation is over the n background pixels. w_i is a weight which should ideally be the inverse of the variance of ρ_i. Assuming that the variance is determined by counting statistics, this gives

$$w_i = 1\big/GE(\rho_i), \qquad (11.2.5.2)$$

where G is the gain of detector, which converts pixel counts to equivalent X-ray photons, and $E(\rho_i)$ is the expectation value of the background counts ρ_i. In practice, the variation in background across the measurement box is usually sufficiently small that all weights can be considered to be equal.

This gives the following equations for a, b and c, as given in Rossmann (1979),

$$\begin{pmatrix} \sum p^2 & \sum pq & \sum p \\ \sum pq & \sum q^2 & \sum q \\ \sum p & \sum q & n \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum p\rho \\ \sum q\rho \\ \sum \rho \end{pmatrix}, \qquad (11.2.5.3)$$

where all summations are over the n background pixels.

11.2.5.1.1. Outlier rejection

It is not unusual for the diffraction pattern to display features other than the Bragg diffraction spots from the crystal of interest. Possible causes are the presence of a satellite crystal or twin component, white-radiation streaks, cosmic rays or zingers. In order to minimize their effect on the determination of the background plane constants, the following outlier rejection algorithm is employed:

1. Determine the background plane constants using a fraction (say 80%) of the background pixels, selecting those with the *lowest* pixel values.
2. Evaluate the fit of all background pixels to this plane, rejecting those that deviate by more than three standard deviations.
3. Re-determine the background plane using all accepted pixels.
4. Re-evaluate the fit of all accepted pixels and reject outliers. If any new outliers are found, re-determine the plane constants.

The rationale for using a subset of the pixels with the lowest pixel values in step (1) is that the presence of zingers or cosmic rays, or a strongly diffracting satellite crystal, can distort the initial calculation of the background plane so much that it becomes difficult to identify the true outliers. Such features will normally only affect a small percentage of the background pixels and will invariably give higher than expected pixel counts. Selecting a subset with the lowest pixel values will facilitate identification of the true outliers. The initial bias in the resulting plane constant c due to this procedure will be corrected in step (3). Poisson statistics are used to evaluate the standard deviations used in outlier rejection, and the standard deviation used in step (2) is increased to allow for the choice of background pixels in step (1).

11.2.5.2. Evaluating the integrated intensity and standard deviation

The summation integration intensity I_s is given by

$$I_s = \sum_{i=1}^{m}(\rho_i - ap_i - bq_i - c), \qquad (11.2.5.4)$$

where the summation is over the m pixels in the peak region of the measurement box. If the peak region has *mm* symmetry, this simplifies to

$$I_s = \sum_{i=1}^{m}(\rho_i - c). \qquad (11.2.5.5)$$

To evaluate the standard deviation, this can be written as

$$I_s = \sum_{i=1}^{m} \rho_i - (m/n) \sum_{j=1}^{n} \rho_j, \qquad (11.2.5.6)$$

where the second summation is over the n background pixels.

The variance in I_s is

$$\sigma_{I_s}^2 = \sum_{i=1}^{m} \sigma_i^2 + (m/n)^2 \sum_{j=1}^{n} \sigma_j^2. \qquad (11.2.5.7)$$

From Poisson statistics this becomes

$$\sigma_{I_s}^2 = \sum_{i=1}^{m} G\rho_i + (m/n)^2 \sum_{j=1}^{n} G\rho_j \qquad (11.2.5.8)$$

$$= G\left[I_s + I_{\rm bg} + (m/n)(m/n) \sum_{j=1}^{n} \rho_j \right], \qquad (11.2.5.9)$$

where $I_{\rm bg}$ is the background summed over all peak pixels. We can also write

$$I_{\rm bg} \simeq (m/n) \sum_{j=1}^{n} \rho_j \qquad (11.2.5.10)$$

(this is only strictly true if the background region has *mm* symmetry). Then

$$\sigma_{I_s}^2 = G\left[I_s + I_{\rm bg} + (m/n) I_{\rm bg} \right]. \qquad (11.2.5.11)$$

This expression shows the importance of the background ($I_{\rm bg}$) in determining the standard deviation of the intensity. For weak reflections, the Bragg intensity (I_s) is often much smaller than the background ($I_{\rm bg}$), and the error in the intensity is determined entirely by the background contribution.

11.2.5.3. *The effect of instrument or detector errors*

Standard-deviation estimates calculated using (11.2.5.11) are generally in quite good agreement with observed differences between the intensities of symmetry-related reflections for weak or medium intensities. This is particularly true if other sources of systematic error are minimized by measuring the *same* reflections five or more times, by doing multiple exposures of the same small oscillation range and then processing the data in space group $P1$. However, even in this latter case, the agreement between strong intensities is significantly worse than that predicted using equation (11.2.5.11). This is consistent with the observation that it is very unusual to obtain merging R factors lower than 0.01, even for very strong reflections where Poisson statistics would suggest merging R factors should be in the range 0.002–0.003.

An experiment in which a diffraction spot recorded on photographic film was scanned many times on an optical microdensitometer showed that the r.m.s. variation in individual pixel values between the scans was greatest for those pixels immediately surrounding the centre of the spot, where the gradient of the optical density was greatest. One explanation for this observation is that these optical densities will be most sensitive to small errors in positioning the reading head, due to vibration or mechanical defects. A simple model for the instrumental contribution to the standard deviation of the spot intensity is obtained by introducing an additional term for each pixel in the spot peak:

$$\sigma_{\rm ins} = K \frac{\delta\rho}{\delta x}, \qquad (11.2.5.12)$$

where $\delta\rho/\delta x$ is the average gradient and K is a proportionality constant. Taking a triangular reflection profile, the gradient and integrated intensity are related by

$$I_s = \frac{1}{12}\left(x^3 + 3x^2 + 5x + 3\right) \frac{\delta\rho}{\delta x}, \qquad (11.2.5.13)$$

where x is the half-width of the reflection (in pixels).

Writing

$$A = \frac{1}{12}\left(x^3 + 3x^2 + 5x + 3\right) \qquad (11.2.5.14)$$

gives

$$\sigma_{\rm ins} = (K/A) I_s, \qquad (11.2.5.15)$$

where the factor A allows for differences in spot size and K is, ideally, a constant for a given instrument.

The total variance in the integrated intensity is then

$$\sigma_{\rm tot}^2 = \sigma_{I_s}^2 + m\sigma_{\rm ins}^2 \qquad (11.2.5.16)$$

$$= G\left[I_s + I_{\rm bg} + (m/n) I_{\rm bg} \right] + m(K/A)^2 I_s^2. \qquad (11.2.5.17)$$

A value for K can be determined by comparing the goodness-of-fit of the standard profiles to individual reflection profiles (of fully recorded reflections) with that calculated from combined Poisson statistics and the instrument error term. Standard deviations estimated using (11.2.5.17) give much more realistic estimates than those based on (11.2.5.11), even for data collected with charge-coupled-device (CCD) detectors where the physical model for the source of the error is clearly not appropriate.

11.2.6. Integration by profile fitting

Providing the background and peak regions are correctly defined, summation integration provides a method for evaluating integrated intensities that is both robust and free from systematic error. For weak reflections, however, many of the pixels in the peak region will contain very little signal (Bragg intensity) but will contribute significantly to the noise because of the Poissonian variation in the background [as shown by the $I_{\rm bg}$ term in equation (11.2.5.11)]. Profile fitting provides a means of improving the signal-to-noise ratio for this class of reflection (but will provide no improvement for reflections where the background level is negligible).

11.2.6.1. *Forming the standard profiles*

In order to apply profile-fitting methods, the first requirement is to derive a 'standard' profile that accurately represents the true reflection profile. Although analytical functions can be used, it is difficult to define a simple function that will cope adequately with the wide variation in spot shapes that can arise in practice. Most programs therefore rely on an empirical profile derived by summing many different spots. The optimum profile is that which provides the best fit to all the contributing reflections, *i.e.* that which minimizes

$$R_2 = \sum_h w_j(h)\left[K_h P_j - \rho_j(h)_{\rm corr} \right]^2, \qquad (11.2.6.1)$$

where P_j is the profile value for the jth pixel, $\rho_j(h)_{\rm corr}$ is the observed background-corrected count at that pixel for reflection h, K_h is a scale factor and $w_j(h)$ is a weight for the jth pixel of reflection h. The summation extends over all reflections contributing to the profile. The weight is given by

$$w_j(h) = 1/\sigma_{hj}^2, \qquad (11.2.6.2)$$

and from Poisson statistics σ_{hj}^2 is the expectation value of the counts at pixel j, and is given by

$$\sigma_{hj}^2 = K_h P_j + \left(a_h p_j + b_h q_j + c_h \right). \qquad (11.2.6.3)$$

After Rossmann (1979), the summation integration intensity $I_s(h)$ can be used to derive a value for K_h:

$$I_s(h) = K_h \sum_{j=1}^{m} P_j. \qquad (11.2.6.4)$$

In equations (11.2.6.3) and (11.2.6.4), as the profile values P_j are not yet determined, a preliminary profile derived, for example, from simple summation of strong reflections used in the detector-parameter refinement can be used, which will give acceptable weights for use in equation (11.2.6.1).

This method of deriving the standard profile is only appropriate for fully recorded reflections. However, in many cases there will be very few or no fully recorded reflections on each image. In such cases the profile is determined by simply adding together the background-corrected pixel counts from all contributing reflections. In the program *MOSFLM* (Leslie, 1992), the profiles are determined using reflections on, typically, ten or more successive images, so that partials will be summed to give the correct fully recorded profile for the majority of the contributing reflections. Tests carried out using standard profiles derived using only fully recorded reflections and equation (11.2.6.1), or using both fully recorded and partially recorded reflections and simple summation, give data of the same quality as judged by the merging statistics.

The reflection profile changes across the face of the detector, due to obliquity of incidence, changes in the projected diffracting volume and geometric factors. In the *MOSFLM* program, this variation is accommodated by determining several standard profiles (typically nine or 25) for different regions of the detector. When evaluating the profile-fitted intensity for a given reflection, a weighted sum of the nearest standard profiles is calculated to provide the best estimate of the true profile at that position on the detector. For the central regions of the detector there will be four contributing profiles, while at the edges there will be between one and three. The weights assigned to each profile vary linearly with the distance from the reflection to the centres of the regions used in determining the standard profiles. An alternative procedure used in *DENZO* (Otwinowski & Minor, 1997) is to evaluate a new profile for each reflection based on spots lying within a pre-specified radius.

11.2.6.2. *Evaluation of the profile-fitted intensity*

Given an appropriate standard profile, the reflection intensity for fully recorded reflections is evaluated by determining the scale factor K and background plane constants a, b, c which minimize

$$R_3 = \sum w_i(KP_i + ap_i + bq_i + c - \rho_i)^2, \qquad (11.2.6.5)$$

where the summation is over all valid pixels in the measurement box. As before,

$$w_i = 1/\sigma_i^2 \qquad (11.2.6.6)$$

and

$$\begin{aligned}\sigma_i^2 &= \text{expectation value of the counts at pixel } i\\ &= ap_i + bq_i + c + JP_i. \qquad (11.2.6.7)\end{aligned}$$

In order to calculate the weights, the background plane constants and summation integration intensity I_s are evaluated as described in Section 11.2.5, at the same time identifying any outliers in the background. The summation integration intensity is used to evaluate the scale factor J in equation (11.2.6.7) using

$$I_s = J \sum_i P_i. \qquad (11.2.6.8)$$

In equation (11.2.6.5), the summation is over all valid pixels within the measurement box. This excludes pixels that are overlapped by neighbouring spots (if any) and any outliers identified in the background region.

Minimizing R_3 with respect to K, a, b and c leads to four linear equations from which K, a, b and c can be determined:

$$\begin{pmatrix} \sum wP^2 & \sum wpP & \sum wqP & \sum wP \\ \sum wpP & \sum wp^2 & \sum wpq & \sum wp \\ \sum wqP & \sum wpq & \sum wq^2 & \sum wq \\ \sum wP & \sum wp & \sum wq & \sum w \end{pmatrix} \begin{pmatrix} K \\ a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \sum wP\rho \\ \sum wp\rho \\ \sum wq\rho \\ \sum w\rho \end{pmatrix}. \qquad (11.2.6.9)$$

The profile-fitted intensity I_p is then given by

$$I_p = K \sum_i P_i. \qquad (11.2.6.10)$$

The standard deviation in the profile-fitted intensity is given by

$$\begin{aligned}\sigma_{I_p}^2 &= \sigma_K^2 \Bigl(\sum_i P_i\Bigr)^2 \qquad (11.2.6.11)\\ &= \Bigl(\sum_i^N w_i\Delta_i^2/(N-4)\Bigr) A_{KK}^{-1} \Bigl(\sum_i P_i\Bigr)^2, \qquad (11.2.6.12)\end{aligned}$$

where

$$\Delta_i = (KP_i + ap_i + bq_i + c - \rho_i), \qquad (11.2.6.13)$$

N is the number of pixels in the summation and A_{KK}^{-1} is the diagonal element for the scale factor K of the inverse normal matrix (used to minimize R_3).

In the case of partially recorded reflections, it is no longer valid to fit the sum of the scaled standard profile and a background plane to all pixels in the measurement box. Partially recorded reflections can have a profile that differs significantly from the standard profile, with the result that the background plane constants take on physically unreasonable values in an attempt to compensate for this difference. Therefore, for partially recorded reflections, the summation in equation (11.2.6.5) is restricted to pixels in the peak region of the measurement box. Minimizing R_3 with respect to the scale factor K then gives

$$\begin{aligned}I_p &= K\sum P_i \qquad (11.2.6.14)\\ &= \Bigl(\sum w_iP_i\rho_i - a\sum w_iP_ip_i - b\sum w_iP_iq_i - c\sum w_iP_i\Bigr)\sum P_i \Big/ \sum w_iP_i^2, \qquad (11.2.6.15)\end{aligned}$$

where all summations are over the peak region only.

It is not possible to derive a standard deviation for partially recorded reflections based on the fit of the scaled standard profile (because partially recorded reflections have a different spot profile). For these reflections, the standard deviation can be calculated using equation (11.2.5.17).

11.2.6.3. *Modifications for very close spots*

In order to apply equation (11.2.6.5), it is necessary to exclude all pixels in the measurement box that are overlapped by a neighbouring spot. This applies not only to the pixels of the reflection being integrated, but also to the pixels of all the reflections used to form the standard profile. Consequently, a pixel should be excluded even if it is only overlapped by a neighbouring spot for one of the reflections used in forming the standard profile. When processing data from large unit cells, this can lead to a very high percentage of the background pixels being rejected and therefore a poor determination of the background plane parameters. In these circumstances, the background plane is determined using only background pixels and excluding only those pixels that are overlapped by neighbours for the reflection actually being integrated. The profile-fitted intensity for both fully recorded and partially recorded reflections is then evaluated in the way described for partially recorded reflections in Section 11.2.6.2, with the summation in equation (11.2.6.15) extending only over peak pixels. The standard deviation in the intensity for partially recorded reflections is derived from equation (11.2.5.17) as before. For fully recorded reflections, the standard deviation has two components: the first is based on the fit of the scaled standard profile to the reflection profile and the second on the contribution from the background:

$$\sigma_I^2 = \sigma_{\rm prof}^2 + \sigma_{\rm bg}^2 \qquad (11.2.6.16)$$

$$= \left[\sum_{i=1}^{m} w_i\Delta_i^2\Big/(m-1)\right]\left[\left(\sum_{i=1}^{m}P_i\right)^2\Big/\sum_{i=1}^{m}w_iP_i^2\right] + (m/n)\sum_{i=1}^{n}(\rho_i - ap_i - bq_i - c)^2, \qquad (11.2.6.17)$$

where m and n are the number of pixels in the peak and background, respectively.

11.2.6.4. *Profile fitting very strong reflections*

For very strong reflections, the background level is very small and equation (11.2.6.15) reduces to

$$I_p \simeq \textstyle\sum w_iP_i\rho_i \sum P_i\Big/\sum w_iP_i^2, \qquad (11.2.6.18)$$

and the weights are given by

$$w_i \simeq 1\big/JP_i. \qquad (11.2.6.19)$$

Substituting for w_i in (11.2.6.18) gives

$$I_p \simeq \textstyle\sum \rho_i. \qquad (11.2.6.20)$$

As pointed out by Z. Otwinowski (personal communication), this shows that for correctly weighted profile fitting, the profile-fitted intensity reduces to the summation integration intensity for very strong intensities.

11.2.6.5. *Profile fitting very weak reflections*

For very weak reflections, all pixels will have very similar counts and therefore all the weights will be the same. For simplicity, consider the case where the profile fit is evaluated only for the peak pixels, then equation (11.2.6.15) reduces to

$$I_P \simeq \textstyle\sum P_i(\rho_i - ap_i - bq_i - c)\sum P_i\Big/\sum P_i^2. \qquad (11.2.6.21)$$

The second and third summations in this equation depend only on the shape of the standard profile. This shows that the intensity is a weighted sum of the individual background-corrected pixel counts (rather than a simple unweighted sum, as is the case for summation integration). Because the values of P_i are a maximum in the centre of the spot, this will place a higher weight on those pixels where the contribution of the Bragg diffraction is greatest, and a very low weight on the peripheral pixels where the Bragg diffraction is weakest. In this way, profile fitting improves the signal-to-noise ratio without the risk of introducing any systematic error that may result by simply reducing the size of the peak region for weak spots.

11.2.6.6. *Improvement provided by profile fitting weak reflections*

For very weak reflections, where all the weights w_i are approximately the same, the variance in I_p using equation (11.2.6.21) is given by

$$\sigma_{I_p}^2 = \textstyle\sum {\rm Var}(\rho_i - ap_i - bq_i - c)P_i^2\left(\sum P_i\Big/\sum P_i^2\right)^2. \qquad (11.2.6.22)$$

Assuming a flat background and very weak intensity, then from Poisson statistics

$${\rm Var}(\rho_i - ap_i - bq_i - c) \simeq G\rho_i, \qquad (11.2.6.23)$$

and as ρ_i has approximately the same value (ρ) for all pixels,

$$\sigma_{I_p}^2 = G\rho\textstyle\sum P_i^2\left(\sum P_i\Big/\sum P_i^2\right)^2 \qquad (11.2.6.24)$$

$$= G\rho\left(\textstyle\sum P_i\right)^2\Big/\textstyle\sum P_i^2. \qquad (11.2.6.25)$$

The variance in the summation integration intensity is simply

$$\sigma_{I_s}^2 = Gm\rho. \qquad (11.2.6.26)$$

The ratio of the variances is thus

$$\sigma_{I_s}^2\big/\sigma_{I_p}^2 = m\textstyle\sum P_i^2\Big/\left(\sum P_i\right)^2. \qquad (11.2.6.27)$$

For a typical spot profile, the right-hand side (which depends only on the shape of the standard profile) has a value of 2, showing that profile fitting can reduce the standard deviation in the integrated intensity by a factor of $(2)^{1/2}$.

11.2.6.7. *Other benefits of profile fitting*

11.2.6.7.1. *Incompletely resolved spots*

If adjacent spots are not fully resolved, there will be a systematic error in the integrated intensity which will be largest for weak spots that are adjacent to very strong spots. However, the profile-fitted intensity will be affected less than the summation integration intensity, because the peripheral pixels (where the influence of neighbouring spots is greatest) are downweighted relative to the central pixels (where the neighbours will have least influence).

Further steps can be taken to minimize the errors caused by overlapping spots. Firstly, when forming the standard profiles, reflections are only included if they are significantly stronger than their nearest neighbours. This will minimize the errors in the standard profiles. Secondly, when evaluating the profile-fitted intensity of a particular reflection, pixels can be omitted if they are adjacent to a pixel that is part of a neighbouring spot (rather than having to be part of that spot).

11.2.6.7.2. *Elimination of peak pixel outliers*

In the same way that outliers in the background region can be identified and rejected (see Section 11.2.5.1.1), it is possible in principle to identify outliers in the peak region of fully recorded reflections as those pixels whose deviation from the scaled standard profile is significantly greater than that expected from counting statistics. This approach works well if the feature that gives rise to the outliers affects only a small fraction of the peak pixels and gives rise to large deviations, and this is the case for some zingers or dead pixels, and for diffraction from small ice crystals when collecting data from cryo-cooled samples.

Another source of outliers is the encroachment of a strong neighbouring spot into the peak region, as discussed in Section 11.2.6.7.1. When dealing with peripheral pixels, the outlier test can be applied to both fully recorded and partially recorded reflections, but a high σ cutoff (*e.g.* 10–20) must be used to avoid rejecting pixels that do not fit the profile simply because they correspond to a partially recorded spot.

11.2.6.7.3. Estimation of overloaded reflections

Owing to the limited dynamic range of current detectors, it is common for many low-resolution spots to contain saturated pixels. Providing the saturation level of the detector is known, such pixels can simply be excluded from the profile fitting, allowing a reasonable estimate of the true intensity (except when the majority of the pixels are saturated). A knowledge of the strong intensities is essential for structure solution based on molecular replacement techniques, and so this is a very useful additional feature of profile fitting.

11.2.6.8. Profile fitting partially recorded reflections

Greenhough & Suddath (1986) have shown that when profile fitting is applied to partially recorded reflections this leads to a systematic error in the individual intensities, but there is no systematic error in the total summed intensity. Although their analysis is strictly only applicable to the case of unweighted profile fitting, experience has shown that even when using weighted profile fitting there is no evidence of systematic errors in the summed profile-fitted intensities of partially recorded reflections. This is particularly important as many data sets collected from frozen crystals have few, if any, fully recorded reflections.

11.2.6.9. Systematic errors in profile-fitted intensities

The fundamental assumption in profile fitting is that the standard profiles accurately reflect the true profile of the reflection being integrated. Errors in the standard profile will result in systematic errors in the profile-fitted intensities. While these errors will often be small compared to the random (Poissonian) error for weak reflections, this is not necessarily the case for strong reflections, as the systematic error is typically a small percentage of the total intensity. Because the standard profiles are derived from the summation of many contributing reflections, small positional errors in spot prediction will lead to a broadening of the standard profile relative to the profile of an individual spot. The same broadening can occur because of the finite sampling interval in the image, which means that a predicted spot position can lie up to half a pixel away from the centre of the measurement box. This error can be minimized by interpolating the pixel values in the image onto a grid which is centred exactly on the predicted position, but the interpolation step itself will inevitably distort the reflection profile. In spite of these difficulties, providing adequate care is taken to determine the crystal and detector parameters accurately (as mentioned in Section 11.2.2), so that the spot positions are predicted to within a small fraction of the overall spot width, there is no suggestion (from merging statistics at least) for significant systematic error, even in the stronger intensities.

I would like to thank Dr A. J. Wonacott, Dr P. Brick and Dr P. R. Evans for many stimulating and critical discussions on all aspects of data integration.

References

Diamond, R. (1969). *Profile analysis in single crystal diffractometry*. *Acta Cryst.* A**25**, 43–55.

Ford, G. C. (1974). *Intensity determination by profile fitting applied to precession photographs*. *J. Appl. Cryst.* **7**, 555–564.

Greenhough, T. J. & Suddath, F. L. (1986). *Oscillation camera data processing. 4. Results and recommendations for the processing of synchrotron radiation data in macromolecular crystallography*. *J. Appl. Cryst.* **19**, 400–409.

Lehmann, M. S. & Larsen, F. K. (1974). *A method for location of the peaks in step-scan-measured Bragg reflexions*. *Acta Cryst.* A**30**, 580–584.

Leslie, A. G. W. (1992). *Recent changes to the MOSFLM package for processing film and image plate data. CCP4 and ESF-EACMB Newsletter on Protein Crystallography*. Warrington: Daresbury Laboratory.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol.* **276**, 307–326.

Rossmann, M. G. (1979). *Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting*. *J. Appl. Cryst.* **12**, 225–238.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *Processing and post-refinement of oscillation camera data*. *J. Appl. Cryst.* **12**, 570–581.

Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *The oscillation method for crystals with very large unit cells*. *Acta Cryst.* A**35**, 901–911.

references

International Tables for Crystallography (2012). Vol. F, Chapter 11.3, pp. 272–281.

Chapter 11.3. Integration, scaling, space-group assignment and post refinement

W. Kabsch

11.3.1. Introduction

The key steps in the processing of diffraction data from single crystals involve (i) modelling of the observed reflection positions in the detector plane, (ii) integration of diffraction intensities, (iii) data correction, scaling and post refinement and (iv) space-group assignment. Much of the theory and many of the methods for carrying out these steps were developed about three decades ago for processing rotation data recorded on film and were subsequently extended in order to fully exploit the capabilities of a variety of electronic area detectors; some CCD (charge-coupled device) and multiwire detectors as well as a new pixel detector specially developed for data collection at synchrotron beamlines allow the recording of finely sliced rotation data because of their fast data read-out. In this article, the principles of the methods are described as employed by the program *XDS* (Chapter 11.6). These apply equally well to rotation images covering small or large oscillation ranges. A large number of other data-processing systems have been developed which differ in the details of the implementations. Some of these packages are described elsewhere in this volume (see Chapters 11.1, 11.2 and 11.4). The theory and practice of processing fine-sliced data have been discussed by Pflugrath (1997).

11.3.2. Modelling rotation images

The observed diffraction pattern, *i.e.* the positions of the reflections recorded on the rotation-data images, is controlled by a small set of parameters which must be accurately determined before integration can start. Approximate values for some of these parameters are given by the experimental setup, whereas others may be completely unknown and must be obtained from the rotation images. This is achieved by the automatic location of strong diffraction spots, the extraction of a primitive lattice basis that yields integer indices for the observed reflections and the subsequent refinement of all parameters to minimize the discrepancies between observed and calculated spot positions in the data images.

11.3.2.1. Coordinate systems and parameters

In the rotation method, the incident-beam wavevector $\mathbf{S}_0$ of length $1/\lambda$ (where λ is the wavelength) is fixed while the crystal is rotated around a fixed axis described by a unit vector $\mathbf{m}_2$. $\mathbf{S}_0$ points from the X-ray source towards the crystal. It is assumed that the incident beam and the rotation axis intersect at one point at which the crystal must be located. This point is defined as the origin of a right-handed orthonormal laboratory coordinate system $\{\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3\}$. This fixed but otherwise arbitrary system is used as a reference frame to specify the setup of the diffraction experiment.

Diffraction data are assumed to be recorded on a fixed planar detector. A right-handed orthonormal detector coordinate system $\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3\}$ is defined such that a point with coordinates X, Y in the detector plane is represented by the vector $(X - X_0)\mathbf{d}_1 + (Y - Y_0)\mathbf{d}_2 + F\mathbf{d}_3$ with respect to the laboratory coordinate system. The origin X_0, Y_0 of the detector plane is found at a distance $|F|$ from the crystal position. It is assumed that the diffraction data are recorded on adjacent non-overlapping rotation images, each covering a constant oscillation range Δ_φ, with image No. 1 starting at spindle angle φ_0.

Diffraction geometry is conveniently expressed with respect to a right-handed orthonormal goniostat system $\{\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3\}$. It is constructed from the rotation axis and the incident-beam direction such that $\mathbf{m}_1 = (\mathbf{m}_2 \times \mathbf{S}_0)/|\mathbf{m}_2 \times \mathbf{S}_0|$ and $\mathbf{m}_3 = \mathbf{m}_1 \times \mathbf{m}_2$. The origin of the goniostat system is defined to coincide with the origin of the laboratory system.

Finally, a right-handed crystal coordinate system $\{\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ and its reciprocal basis $\{\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*\}$ are defined to represent the unrotated crystal, *i.e.* at rotation angle $\varphi = 0°$, such that any reciprocal-lattice vector can be expressed as $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$, where h, k, l are integers.

As shown in Section 11.3.2.2, the location of all diffraction peaks recorded in the data images can be computed from the parameters $\mathbf{S}_0$, $\mathbf{m}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, X_0, Y_0, F, $\mathbf{d}_1$, $\mathbf{d}_2$, $\mathbf{d}_3$, φ_0 and Δ_φ. In addition, knowledge of the shape and extent of the diffraction spots is required for accurate estimations of their intensities. This can be achieved by a Gaussian model involving two parameters: the standard deviations of the reflecting range, σ_M, and of the beam divergence, σ_D (see Section 11.3.2.3). This leads to an integration region around the spot defined by the parameters δ_M and δ_D, which are typically chosen to be 6–10 times larger than σ_M and σ_D, respectively.

11.3.2.2. Spot prediction

Let $\mathbf{p}_0^*$ denote any arbitrary reciprocal-lattice vector if the crystal has not been rotated, *i.e.* at rotation angle $\varphi = 0°$. Depending on the diffraction geometry, $\mathbf{p}_0^*$ may be rotated into a position fulfilling the reflecting condition. The required rotation angle φ and the coordinates X, Y of the diffracted beam at its intersection with the detector plane can be found from $\mathbf{p}_0^*$ as follows.

$\mathbf{p}_0^*$ can be expressed by its components with respect to the orthonormal goniostat system as

$$\mathbf{p}_0^* = \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^*).$$

Rotation by φ around axis $\mathbf{m}_2$ changes $\mathbf{p}_0^*$ into $\mathbf{p}^*$,

$$\begin{aligned}\mathbf{p}^* &= D(\mathbf{m}_2, \varphi)\mathbf{p}_0^* \\ &= \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*) + [\mathbf{p}_0^* - \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}_0^*)]\cos\varphi + \mathbf{m}_2 \times \mathbf{p}_0^* \sin\varphi \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}_0^* \cos\varphi + \mathbf{m}_3 \cdot \mathbf{p}_0^* \sin\varphi) + \mathbf{m}_2\mathbf{m}_2 \cdot \mathbf{p}_0^* \\ &\quad + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}_0^* \cos\varphi - \mathbf{m}_1 \cdot \mathbf{p}_0^* \sin\varphi) \\ &= \mathbf{m}_1(\mathbf{m}_1 \cdot \mathbf{p}^*) + \mathbf{m}_2(\mathbf{m}_2 \cdot \mathbf{p}^*) + \mathbf{m}_3(\mathbf{m}_3 \cdot \mathbf{p}^*).\end{aligned}$$

The incident-beam and diffracted-beam wavevectors, $\mathbf{S}_0$ and $\mathbf{S}$, have their termini on the Ewald sphere and satisfy the Laue equations

$$\mathbf{S} = \mathbf{S}_0 + \mathbf{p}^*, \quad \mathbf{S}^2 = \mathbf{S}_0^2 \quad \Longrightarrow \quad \mathbf{p}^{*2} = -2\mathbf{S}_0 \cdot \mathbf{p}^* = \mathbf{p}_0^{*2}.$$

If $\rho = [\mathbf{p}_0^{*2} - (\mathbf{p}_0^* \cdot \mathbf{m}_2)^2]^{1/2}$ denotes the distance of $\mathbf{p}_0^*$ from the rotation axis, solutions for $\mathbf{p}^*$ and φ can be obtained in terms of $\mathbf{p}_0^*$ as

Copyright © 2012 International Union of Crystallography

$$\begin{aligned}
&\mathbf{p}^* \cdot \mathbf{m}_3 = [-\mathbf{p}_0^{*2}/2 - (\mathbf{p}_0^* \cdot \mathbf{m}_2)(\mathbf{S}_0 \cdot \mathbf{m}_2)]/\mathbf{S}_0 \cdot \mathbf{m}_3\\
&\mathbf{p}^* \cdot \mathbf{m}_2 = \mathbf{p}_0^* \cdot \mathbf{m}_2\\
&\mathbf{p}^* \cdot \mathbf{m}_1 = \pm[\rho^2 - (\mathbf{p}^* \cdot \mathbf{m}_3)^2]^{1/2}\\
&\quad \cos\varphi = [(\mathbf{p}^* \cdot \mathbf{m}_1)(\mathbf{p}_0^* \cdot \mathbf{m}_1) + (\mathbf{p}^* \cdot \mathbf{m}_3)(\mathbf{p}_0^* \cdot \mathbf{m}_3)]/\rho^2\\
&\quad \sin\varphi = [(\mathbf{p}^* \cdot \mathbf{m}_1)(\mathbf{p}_0^* \cdot \mathbf{m}_3) - (\mathbf{p}^* \cdot \mathbf{m}_3)(\mathbf{p}_0^* \cdot \mathbf{m}_1)]/\rho^2.
\end{aligned}$$

In general there are two solutions according to the sign of $\mathbf{p}^* \cdot \mathbf{m}_1$. If $\rho^2 < (\mathbf{p}^* \cdot \mathbf{m}_3)^2$ or $\mathbf{p}_0^{*2} > 4\mathbf{S}_0^2$ the Laue equations have no solution and the reciprocal-lattice point $\mathbf{p}_0^*$ is in the 'blind' region.

If $F\mathbf{S} \cdot \mathbf{d}_3 > 0$ the diffracted beam intersects the detector plane at the point

$$\begin{aligned}
F\mathbf{S}/\mathbf{S} \cdot \mathbf{d}_3 &= (F\mathbf{S} \cdot \mathbf{d}_1/\mathbf{S} \cdot \mathbf{d}_3)\mathbf{d}_1 + (F\mathbf{S} \cdot \mathbf{d}_2/\mathbf{S} \cdot \mathbf{d}_3)\mathbf{d}_2 + F\mathbf{d}_3\\
&= (X - X_0)\mathbf{d}_1 + (Y - Y_0)\mathbf{d}_2 + F\mathbf{d}_3,
\end{aligned}$$

which leads to a diffraction spot recorded at detector coordinates

$$\begin{aligned}
X &= X_0 + F\mathbf{S} \cdot \mathbf{d}_1/\mathbf{S} \cdot \mathbf{d}_3,\\
Y &= Y_0 + F\mathbf{S} \cdot \mathbf{d}_2/\mathbf{S} \cdot \mathbf{d}_3.
\end{aligned}$$

11.3.2.3. Standard spot shape

A reciprocal-lattice point crosses the Ewald sphere by the shortest route only if the crystal happens to be rotated about an axis perpendicular to both the diffracted-beam and incident-beam wavevectors, the 'β axis' $\mathbf{e}_1 = \mathbf{S} \times \mathbf{S}_0/|\mathbf{S} \times \mathbf{S}_0|$, as introduced by Schutt & Winkler (1977). Rotation around the fixed axis $\mathbf{m}_2$, as enforced by the rotation camera, thus leads to an increase in the length of the shortest path by the factor $1/|\mathbf{e}_1 \cdot \mathbf{m}_2|$. This motivated the introduction of a coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, specific for each reflection, which has its origin on the surface of the Ewald sphere at the terminus of the diffracted beam wavevector $\mathbf{S}$,

$$\begin{aligned}
\mathbf{e}_1 &= \mathbf{S} \times \mathbf{S}_0/|\mathbf{S} \times \mathbf{S}_0|,\\
\mathbf{e}_2 &= \mathbf{S} \times \mathbf{e}_1/|\mathbf{S} \times \mathbf{e}_1|,\\
\mathbf{e}_3 &= (\mathbf{S} + \mathbf{S}_0)/|\mathbf{S} + \mathbf{S}_0|.
\end{aligned}$$

The unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are tangential to the Ewald sphere, while $\mathbf{e}_3$ is perpendicular to $\mathbf{e}_1$ and $\mathbf{p}^* = \mathbf{S} - \mathbf{S}_0$. The shape of a reflection, as represented with respect to $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, then no longer contains geometrical distortions resulting from the fixed rotation axis of the camera and the oblique incidence of the diffracted beam on a flat detector. Instead, all reflections appear as if they had followed the shortest path through the Ewald sphere and had been recorded on the surface of the sphere.

A detector pixel at X', Y' in the neighbourhood of the reflection centre X, Y, when the crystal is rotated by φ' instead of φ, is mapped to the profile coordinates ε_1, ε_2, ε_3 by the following procedure:

$$\begin{aligned}
&\mathbf{S}' = \frac{(X' - X_0)\mathbf{d}_1 + (Y' - Y_0)\mathbf{d}_2 + F\mathbf{d}_3}{\lambda \cdot [(X' - X_0)^2 + (Y' - Y_0)^2 + F^2)]^{1/2}}\\
&\varepsilon_1 = \mathbf{e}_1 \cdot (\mathbf{S}' - \mathbf{S})180/(|\mathbf{S}|\pi)\\
&\varepsilon_2 = \mathbf{e}_2 \cdot (\mathbf{S}' - \mathbf{S})180/(|\mathbf{S}|\pi)\\
&\varepsilon_3 = \mathbf{e}_3 \cdot [D(\mathbf{m}_2, \varphi' - \varphi)\mathbf{p}^* - \mathbf{p}^*]180/(|\mathbf{p}^*|\pi) \simeq \zeta \cdot (\varphi' - \varphi)\\
&\quad \zeta = \mathbf{m}_2 \cdot \mathbf{e}_1.
\end{aligned}$$

ζ corrects for the increased path length of the reflection through the Ewald sphere and is closely related to the reciprocal Lorentz correction factor

$$L^{-1} = \frac{|\mathbf{m}_2 \cdot (\mathbf{S} \times \mathbf{S}_0)|}{(|\mathbf{S}| \cdot |\mathbf{S}_0|)} = \left|\zeta \cdot \sin\angle(\mathbf{S}, \mathbf{S}_0)\right|.$$

Because of crystal mosaicity and beam divergence, the intensity of a reflection is smeared around the diffraction maximum. The fraction of total reflection intensity found in the volume element $\mathrm{d}\varepsilon_1\mathrm{d}\varepsilon_2\mathrm{d}\varepsilon_3$ at ε_1, ε_2, ε_3 can be approximated by Gaussian functions:

$$\begin{aligned}
&\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)\mathrm{d}\varepsilon_1\mathrm{d}\varepsilon_2\mathrm{d}\varepsilon_3\\
&\quad = \frac{\exp(-\varepsilon_1^2/2\sigma_D^2)}{(2\pi)^{1/2}\sigma_D}\,\mathrm{d}\varepsilon_1 \times \frac{\exp(-\varepsilon_2^2/2\sigma_D^2)}{(2\pi)^{1/2}\sigma_D}\,\mathrm{d}\varepsilon_2 \times \frac{\exp(-\varepsilon_3^2/2\sigma_M^2)}{(2\pi)^{1/2}\sigma_M}\,\mathrm{d}\varepsilon_3.
\end{aligned}$$

11.3.2.4. Spot centroids and partiality

The intensity of a reflection can be completely recorded on one image or distributed among several adjacent images. The fraction R_j of total intensity recorded on image j, the 'partiality' of the reflection, can be derived from the distribution function $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ as

$$\begin{aligned}
R_j &= \int\limits_{-\infty}^{\infty} \mathrm{d}\varepsilon_1 \int\limits_{-\infty}^{\infty} \mathrm{d}\varepsilon_2 \int\limits_{\zeta[\varphi_0+(j-1)\Delta_\varphi-\varphi]}^{\zeta(\varphi_0+j\Delta_\varphi-\varphi)} \mathrm{d}\varepsilon_3\, \omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)\\
&= \frac{1}{(2\pi)^{1/2}\sigma_M/|\zeta|} \int\limits_{\varphi_0+(j-1)\Delta_\varphi}^{\varphi_0+j\Delta_\varphi} \exp[-(\varphi' - \varphi)^2/2(\sigma_M/|\zeta|)^2]\,\mathrm{d}\varphi'\\
&= \big(\mathrm{erf}[|\zeta|(\varphi_0 + j\Delta_\varphi - \varphi)/(2)^{1/2}\sigma_M]\\
&\quad - \mathrm{erf}\{|\zeta|[\varphi_0 + (j-1)\Delta_\varphi - \varphi]/(2)^{1/2}\sigma_M\}\big)/2.
\end{aligned}$$

The integral is evaluated by using a numerical approximation of the error function, erf (Abramowitz & Stegun, 1972).

While the spot centroids in the detector plane are usually good estimates for the detector position of the diffraction maximum, the angular centroid about the rotation axis,

$$Z = \varphi_0 + \Delta_\varphi \cdot \sum_{j=-\infty}^{\infty} (j - \tfrac{1}{2})R_j \simeq \varphi,$$

can be a rather poor guess for the true φ angle of the maximum. Its accuracy depends strongly on the value of φ and the size of the oscillation range $\Delta\varphi$ relative to the mosaicity σ_M of the crystal. For a reflection fully recorded on image j, the value $Z = \varphi_0 + (j - \frac{1}{2}) \cdot \Delta_\varphi$ will always be obtained, which is correct only if φ accidentally happens to be close to the centre of the rotation range of the image. In contrast, the φ angle of a partial reflection recorded on images j and $j + 1$ is closely approximated by $Z = \varphi_0 + [j + (R_{j+1} - R_j)/2] \cdot \Delta_\varphi$. If many images contribute to the spot intensity, $Z(\varphi)$ is always an excellent approximation to the ideal angular position φ when the Laue equations are satisfied; in fact, in the limiting case of infinitely fine-sliced data it can be shown that $\lim_{\Delta\varphi\to 0} Z(\varphi) = \varphi$.

Most refinement routines minimize the discrepancies between the predicted φ angles and their approximations obtained from the observed Z centroids and must therefore carefully distinguish between fully and partially recorded reflections. However, this distinction is unnecessary if the observed Z centroids are instead compared with their analytic forms, because the sensitivity of the centroid positions to the diffraction parameters is correctly weighted in either case (see Section 11.3.2.8).

11.3.2.5. Localizing diffraction spots

Often, some of the parameters controlling the diffraction experiment are either completely unknown or available only at a crude approximation. Accurate values for all parameters must be obtained from the recorded data, *i.e.* from a list of the coordinates of strong spots occurring in the images. As implemented in *XDS*, this list is obtained from all or a subset of the data images by the following procedure. Firstly, each pixel value is compared with the mean value and standard deviation of surrounding pixels in the same image and classified as a strong pixel if its value exceeds the mean by a given multiple (typically 3–5) of the standard deviation. Values of the strong pixels and their location addresses and image running numbers are saved in a file. After the scan, a hash table of sufficient size is allocated to accommodate the strong pixels from the file together with their addresses (for a discussion of the hash technique, see Wirth, 1976). As several strong pixels may belong to the same spot, they are labelled with a unique spot number so that any two such pixels which can be connected by direct strong neighbours in two or three dimensions (if there are adjacent images) belong to the same spot (equivalence class). The labelling is achieved by the highly efficient algorithm for the recording of equivalence classes developed by Rem (see Dijkstra, 1976). On termination, a list X_i', Y_i', Z_i' ($i = 1, \ldots, n$) of the centroids of n strong spots is available.

11.3.2.6. Basis extraction

Any reciprocal-lattice vector can be written in the form $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$, where h, k, l are integers and $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are reciprocal basis vectors of the lattice. The basis vectors which describe the orientation, metric and symmetry of the crystal, as well as the reflection indices h, k, l, have to be determined from the list of strong diffraction spots X_i', Y_i', Z_i' ($i = 1, \ldots, n$). Ideally, each spot corresponds to a reciprocal-lattice vector $\mathbf{p}_0^*$ which satisfies the Laue equations after a crystal rotation by φ. Substituting the observed value Z' for the unknown φ angle (see Section 11.3.2.4), $\mathbf{p}_0^*$ is found from the observed spot coordinates as

$$\begin{aligned}\mathbf{p}_0^* &= D(\mathbf{m}_2, -Z')(\mathbf{S}' - \mathbf{S}_0)\\ \mathbf{S}' &= \frac{(X' - X_0)\mathbf{d}_1 + (Y' - Y_0)\mathbf{d}_2 + F\mathbf{d}_3}{\lambda \cdot [(X' - X_0)^2 + (Y' - Y_0)^2 + F^2]^{1/2}}.\end{aligned}$$

Unfortunately, the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$) derived from the above list of strong diffraction spots often contain a number of 'aliens' (spots arising from fluctuations in the background, from ice or from satellite crystals) and a robust method has to be used which is still capable of recognizing the dominant lattice. One approach, suggested by Bricogne (1986) and implemented in a number of variants (Otwinowski & Minor, 1997; Steller *et al.*, 1997), is to identify a lattice basis as the three shortest linear independent vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$, each at a maximum of the Fourier transform $\sum_{i=1}^{n} \cos(2\pi \mathbf{b} \cdot \mathbf{p}_{0i}^*)$. Alternatively, a reciprocal basis for the dominant lattice can be determined from short differences between the reciprocal-lattice vectors (Howard, 1986; Kabsch, 1988*a*). As implemented in *XDS*, a lattice basis is found by the following procedure.

The list of given reciprocal-lattice points $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$) is first reduced to a small number m of low-resolution difference-vector clusters $\mathbf{v}_\mu^*$ ($\mu = 1, \ldots, m$). f_μ is the population of a difference-vector cluster $\mathbf{v}_\mu^*$; that is, the number of times the difference between any two reciprocal-lattice vectors $\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*$ is approximately equal to $\mathbf{v}_\mu^*$. In a second step, three linear independent vectors $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are selected among all possible triplets of difference-vector clusters that maximize the function Q,

$$\begin{aligned}Q(\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*) &= \sum_{\mu=1}^{m} f_\mu q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu),\\ q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu) &= \exp\big(-2\sum_{k=1}^{3}\{[\max(|\xi_k^\mu - h_k^\mu| - \varepsilon, 0)/\varepsilon]^2\\ &\qquad + [\max(|h_k^\mu| - \delta, 0)]^2\}\big),\end{aligned}$$

where

$$\xi_k^\mu = \mathbf{v}_\mu^* \cdot \mathbf{b}_k, \quad \mathbf{v}_\mu^* = \sum_{k=1}^{3} \xi_k^\mu \mathbf{b}_k^*, \quad \mathbf{b}_k \cdot \mathbf{b}_l^* = \begin{cases} 1 & \text{if } k = l\\ 0 & \text{otherwise}\end{cases}$$

and h_k^μ is the nearest integer to ξ_k^μ. The absolute maximum of Q is assumed if all difference vectors can be expressed as small integral multiples of the best triplet. Deviations from this ideal situation are quantified by the quality measure q. The value of q declines sharply if the expansion coefficients ξ_k^μ deviate by more than ε from their nearest integers h_k^μ or if the indices are absolutely larger than δ. The constraint on the allowed range of indices prevents the selection of a spurious triplet of very short difference-vector clusters which might be present in the set. Excellent results have been obtained using $\varepsilon = 0.05$ and $\delta = 5$. The best vector triplet thus found is refined against the observed difference-vector clusters. Finally, a reduced cell is derived from the refined reciprocal-base vector triplet (see Section 11.3.6).

11.3.2.7. Indexing

Once a basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ of the lattice is available, integral indices h_i, k_i, l_i must be assigned to each reciprocal-lattice vector $\mathbf{p}_{0i}^*$ ($i = 1, \ldots, n$). Using the integers nearest to $\mathbf{p}_{0i}^* \cdot \mathbf{b}_k$ ($k = 1, 2, 3$) as indices of the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ could easily lead to a misindexing of longer vectors because of inaccuracies in the basis vectors $\mathbf{b}_k$ and the initial values of the parameters describing the instrumental setup. A more robust solution of the indexing problem is provided by the local indexing method, which assigns only small index differences $h_i - h_j, k_i - k_j, l_i - l_j$ between pairs of neighbouring reciprocal-lattice vectors (Kabsch, 1993).

The reciprocal-lattice points can be considered as nodes of a tree. The tree connects the n points to each other with the connections as its branches. The length ℓ_{ij} of a possible branch between nodes i and j is defined here as

$$\begin{aligned}\ell_{ij} = 1 - \exp\big(-2\sum_{k=1}^{3}\{&[\max(|\xi_k^{ij} - h_k^{ij}| - \varepsilon, 0)/\varepsilon]^2\\ &+ [\max(|h_k^{ij}| - \delta, 0)]^2\}\big),\end{aligned}$$

where

$$\xi_k^{ij} = (\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*) \cdot \mathbf{b}_k,$$

h_k^{ij} is the nearest integer of ξ_k^{ij} and $k = 1, 2, 3$. Reliable index differences are indicated by short branches; in fact, ℓ_{ij} is 0 if none of the indices h_k^{ij} is absolutely larger than δ and the ξ_k^{ij} are integer values to within ε. Typical values are $\varepsilon = 0.05$ and $\delta = 5$. Defining the length of a tree as the sum of the lengths of its branches, a shortest tree among all n^{n-2} possible trees is determined using the elegant algorithm described by Dijkstra (1976). Starting with arbitrary indices 0, 0, 0 for the root node, the local indexing method then consists of traversing the shortest tree and thereby assigning each node the indices of its predecessor plus the small index differences between the two nodes.

During traversal of the tree, each node is also given a subtree number. Starting with subtree number 1 for the root node, each successor node is given the same subtree number as its predecessor if the length of the connecting branch is below a minimal length $\ell_{\min}$. Otherwise, its subtree number is incremented by 1. Thus, all nodes in the same subtree have internally consistent reflection indices. Defining the size of a subtree by the number of its nodes, 'aliens' are usually found in small subtrees. Finally, a constant index offset is determined such that the centroids of the observed reciprocal-lattice points $\mathbf{p}_{0i}^*$ belonging to the largest subtree and their corresponding grid vectors $\sum_{k=1}^{3} h_k^i \mathbf{b}_k^*$ are as close as possible. This offset is added to the indices of each reciprocal-lattice point.

11.3.2.8. Refinement

For a fixed detector, the diffraction pattern depends on the parameters $\mathbf{S}_0$, $\mathbf{m}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, X_0, Y_0, F. Starting values for the parameters can be obtained by the procedures described above, which do not rely on prior knowledge of the crystal orientation, space-group symmetry or unit-cell metric. Better estimates of the parameter values, as required for the subsequent integration step, can be obtained by the method of least squares from the list of n observed indexed reflection centroids h_i, k_i, l_i, X_i', Y_i', Z_i' ($i = 1, \ldots, n$). In this method, the parameters are chosen to minimize a weighted sum of squares of the residuals

$$E = w_X \sum_{i=1}^{n} (\Delta_X^i)^2 + w_Y \sum_{i=1}^{n} (\Delta_Y^i)^2 + w_Z \sum_{i=1}^{n} (\Delta_Z^i)^2.$$

The residuals between the calculated (X_i, Y_i, Z_i) and observed spot centroids are

$$\begin{aligned}
\Delta_X^i &= X_i - X_i' = X_0 + F\mathbf{S}_i \cdot \mathbf{d}_1/\mathbf{S}_i \cdot \mathbf{d}_3 - X_i' \\
\Delta_Y^i &= Y_i - Y_i' = Y_0 + F\mathbf{S}_i \cdot \mathbf{d}_2/\mathbf{S}_i \cdot \mathbf{d}_3 - Y_i' \\
\Delta_Z^i &= Z_i - Z_i' = \varphi_0 + \Delta_\varphi \sum_{j=-\infty}^{\infty} (j - \tfrac{1}{2}) R_j^i - Z_i'.
\end{aligned}$$

Let s_μ ($\mu = 1, \ldots, k$) denote the k independent parameters for which initial estimates are available. Expanding the residuals to first order in the parameter changes δs_μ gives

$$\Delta(s_\mu + \delta s_\mu) \simeq \Delta(s_\mu) + \sum_{\mu=1}^{k} \frac{\partial \Delta}{\partial s_\mu} \delta s_\mu.$$

The parameters should be changed in such a way as to minimize $E(\delta s_\mu)$, which implies $\partial E/\partial \delta s_\mu = 0$ for $\mu = 1, \ldots, k$. The δs_μ are found as the solution of the k normal equations

$$\begin{aligned}
&\sum_{\mu'=1}^{k} \left(w_X \sum_{i=1}^{n} \frac{\partial \Delta_X^i}{\partial s_\mu} \frac{\partial \Delta_X^i}{\partial s_{\mu'}} + w_Y \sum_{i=1}^{n} \frac{\partial \Delta_Y^i}{\partial s_\mu} \frac{\partial \Delta_Y^i}{\partial s_{\mu'}} + w_Z \sum_{i=1}^{n} \frac{\partial \Delta_Z^i}{\partial s_\mu} \frac{\partial \Delta_Z^i}{\partial s_{\mu'}} \right) \delta s_{\mu'} \\
&\quad = -\left(w_X \sum_{i=1}^{n} \Delta_X^i \frac{\partial \Delta_X^i}{\partial s_\mu} + w_Y \sum_{i=1}^{n} \Delta_Y^i \frac{\partial \Delta_Y^i}{\partial s_\mu} + w_Z \sum_{i=1}^{n} \Delta_Z^i \frac{\partial \Delta_Z^i}{\partial s_\mu} \right).
\end{aligned}$$

The parameters are corrected by δs_μ and a new cycle of refinement is started until a minimum of E is reached. The weights

$$w_X = 1/\sum_{i=1}^{n} (\Delta_X^i)^2, \quad w_Y = 1/\sum_{i=1}^{n} (\Delta_Y^i)^2, \quad w_Z = 1/\sum_{i=1}^{n} (\Delta_Z^i)^2$$

are calculated with the current guess for s_μ at the beginning of each cycle.

The derivatives appearing in the normal equations can be worked out from the definitions given in Sections 11.3.2.2 and 11.3.2.4 and only the form of the gradient of the Z residuals is shown. Assuming $\sigma_i = \sigma_M/|\zeta_i|$ ($i = 1, \ldots, n$) is constant for each reflection, the gradients of the Z residuals are obtained from the chain rule and the relation $\mathrm{d}\,\mathrm{erf}(z)/\mathrm{d}z = (2/\pi^{1/2})\exp(-z^2)$.

$$\begin{aligned}
\frac{\partial \Delta_Z^i}{\partial s_\mu} &= \frac{\partial \Delta_Z^i}{\partial \varphi_i} \frac{\partial \varphi_i}{\partial s_\mu} \\
\frac{\partial \Delta_Z^i}{\partial \varphi_i} &= \frac{\Delta_\varphi}{(2\pi)^{1/2} \sigma_i} \sum_{j=-\infty}^{\infty} \exp[-(\varphi_0 + j\Delta_\varphi - \varphi_i)^2 / 2\sigma_i^2] \\
\frac{\partial \varphi_i}{\partial s_\mu} &= \cos \varphi_i \frac{\partial \sin \varphi_i}{\partial s_\mu} - \sin \varphi_i \frac{\partial \cos \varphi_i}{\partial s_\mu}.
\end{aligned}$$

Obviously, $\partial \Delta_Z^i / \partial s_\mu$ is small for a fully recorded reflection because of the small values of all exponentials appearing in $\partial \Delta_Z^i / \partial \varphi_i$. In contrast, the gradient for a partial reflection that is equally recorded on two adjacent images is most sensitive to parameter variations because one of the exponentials assumes its maximum value. In the limiting case of infinitely fine-sliced data it can be shown that $\lim_{\Delta\varphi \to 0} \partial \Delta_Z^i / \partial \varphi_i = 1$. Thus, the refinement scheme based on observed Z centroids, as described here and implemented in *XDS*, is applicable to fine-sliced data and also to data recorded with a large oscillation range.

11.3.3. Integration

Assuming that the diffraction parameters have been refined successfully as described above, the intensity of a reflection is distributed in the neighbourhood of the predicted location of the diffraction peak among detector pixels of one or several adjacent rotation images. Accurate integration requires several steps: determination of a reflection mask, estimation of the background, generation of reference profiles and integration by profile fitting.

The intensity distribution of a reflection can be modelled analytically or derived from the observed profiles of neighbouring strong spots. For the rotation method, the profile shape depends strongly on the specific path of the reflection through the Ewald sphere and on variations in the angle of incidence of the diffracted beam on a flat detector. These geometrical distortions can be eliminated by mapping the reflections onto the coordinate system defined in Section 11.3.2.3, which simplifies the task of modelling the expected intensity distribution as all reflection profiles become similar.

11.3.3.1. Reflection mask

The parameters σ_M and σ_D of the Gaussian model (see Section 11.3.2.3) used to describe reflection shape can be determined automatically from one or more data images by the following procedure.

(i) Identify and mark strong pixels in the data image.
(ii) Assign the indices of the nearest reflection to each strong pixel.
(iii) Sort the strong pixels by the assigned reflection indices such that pixels with the same indices follow each other in the list.
(iv) For each strong reflection find the rectangular box that encloses all of the strong pixels belonging to the reflection.
(v) Increase the box slightly and use all pixels within the box that are not strong for background determination.
(vi) Subtract the background and determine the centroid and variance s^2 of the intensity-weighted diffracted-beam directions $\lambda\mathbf{S}'$ associated with each strong pixel belonging to the spot (see Section 11.3.2.3).
(vii) Reject the spot if the centroid position deviates too much from the calculated spot location.
(viii) Calculate φ and ζ for the accepted reflection and save the three values φ, ζ and s^2 in a list.

The standard deviation of the beam divergence is obtained directly from this list of n reflections as

$$\sigma_D^2 = \frac{1}{n} \sum_{j=1}^{n} s_j^2.$$

Determination of the standard deviation of the reflecting range, the mosaicity σ_M, requires additional considerations. For each of the n reflections from the list above, let τ denote the angular difference between the rotation angle φ at its Bragg maximum and the centre of the oscillation angles covered by the image. The fraction of observed reflection intensity is (see Section 11.3.2.4)

$$R(\tau;\sigma_M/\zeta)=\frac{1}{2\Delta_\varphi}\left[\operatorname{erf}\left(\frac{\tau+\Delta_\varphi/2}{2^{1/2}\sigma_M/\zeta}\right)-\operatorname{erf}\left(\frac{\tau-\Delta_\varphi/2}{2^{1/2}\sigma_M/\zeta}\right)\right].$$

For a given σ_M/ζ the function $R(\tau; \sigma_M/\zeta)$ assumes its maximum at $\tau = 0$ and declines as $|\tau|$ increases. The decline depends strongly on the mosaicity σ_M and on the path length of the reflection through the Ewald sphere, which is accounted for by the factor $1/\zeta$. For a large mosaicity $R(\tau; \sigma_M/\zeta)$ declines slowly, which explains why for such crystals many reflections with large $|\tau|$ values can be observed on a data image. Clearly, the list of strong spots located by the automatic procedure described above contains information about the mosaicity of the crystal. The problem of finding σ_M from this list can be solved if one considers each τ value as a random variable drawn from a probability distribution $R(\tau; \sigma_M/\zeta)$ with population parameter σ_M/ζ. The mosaicity σ_M can then be estimated so that it maximizes the likelihood (joint probability)

$$L(\sigma_M)=R(\tau_1;\sigma_M/\zeta_1)\cdot R(\tau_2;\sigma_M/\zeta_2)\ldots R(\tau_n;\sigma_M/\zeta_n).$$

The parameters σ_D and σ_M are mainly used to specify the integration region around the spot defined by the parameters δ_M and δ_D, which are typically chosen to be 6–10 times larger than σ_M and σ_D, respectively (see Section 11.3.2.1). The reflection mask thus comprises all image pixels that satisfy

$$|\varepsilon_1|\le\delta_D/2,\quad |\varepsilon_2|\le\delta_D/2,\quad |\varepsilon_3|\le\delta_M/2$$

when mapped to the profile coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ defined in Section 11.3.2.3. In addition, pixels are excluded from the mask if they are closer to the predicted Bragg peak of an intruding reflection from the neighbourhood.

11.3.3.2. Background

The region around a spot is assumed to have been chosen to be large enough to include a sufficient number of pixels which can be used for determination of the background. Background determination, as implemented in *XDS*, begins by sorting all pixels belonging to a reflection by increasing intensity. For weak or absent reflections, these values should represent a random sample drawn from a normal distribution. If this is not the case, the pixel with the largest intensity is removed until the sampling distribution of the remaining smaller items satisfies the expected distribution. This method will also exclude pixels with unexpected high values, such as ice reflections. The background, determined as the mean value of the accepted pixels, is systematically overestimated for strong spots because of some residual intensity extending into the accepted background pixels. This residual intensity is estimated from the expected distribution $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ defined in Section 11.3.2.3 and removed from the final background value.

11.3.3.3. Standard profiles

Reflection profiles are represented on the Ewald sphere within a domain D_0 comprising $2n_1 + 1$, $2n_2 + 1$ and $2n_3 + 1$ equidistant gridpoints along $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$, respectively. The sampling distances between adjacent grid points are then $\Delta_1 = \delta_D/(2n_1 + 1)$, $\Delta_2 = \delta_D/(2n_2 + 1)$ and $\Delta_3 = \delta_M/(2n_3 + 1)$. Thus, grid coordinate ν_3 ($\nu_3 = -n_3, \ldots, n_3$) covers the set of rotation angles

$$\Gamma_{\nu_3}=\{\varphi'|(\nu_3-\tfrac{1}{2})\Delta_3\le(\varphi'-\varphi)\cdot\zeta\le(\nu_3+\tfrac{1}{2})\Delta_3\}.$$

Contributions to the spot intensity come from one or several adjacent data images ($j = j_1, \ldots, j_2$), each covering the set of rotation angles

$$\Gamma_j=\{\varphi'|\varphi_0+(j-1)\Delta_\varphi\le\varphi'\le\varphi_0+j\Delta_\varphi\}.$$

Assuming Gaussian profiles along $\mathbf{e}_3$ for all reflections (see Section 11.3.2.3), the fraction of counts (after subtraction of the background) contributed by data frame j to grid coordinate ν_3 is

$$f_{\nu_3 j}\simeq\frac{\int_{\Gamma_j\cap\Gamma_{\nu_3}}\exp[-(\varphi'-\varphi)^2/2\sigma^2]\,\mathrm{d}\varphi'}{\int_{\Gamma_j}\exp[-(\varphi'-\varphi)^2/2\sigma^2]\,\mathrm{d}\varphi',}$$

where $\sigma = \sigma_M/|\zeta|$. The integrals can be expressed in terms of the error function, for which efficient numerical approximations are available (Abramowitz & Stegun, 1972). Finally, each pixel in data image j belonging to the reflection is subdivided into 5 × 5 areas of equal size and $f_{\nu_3 j}/25$ of the pixel signal is added to the profile value at grid coordinates ν_1, ν_3, ν_3 corresponding to each subdivision.

This complicated procedure leads to more uniform intensity profiles for all reflections than using their untransformed shape. This simplifies the task of modelling the expected intensity distribution needed for integration by profile fitting. As implemented in *XDS*, reference profiles are learnt every 5° of crystal rotation at nine positions on the detector, each covering an equal area of the detector face. In the learning phase, profile boxes of the strong reflections are normalized and added to their nearest reference profile boxes. The contributions are weighted according to the distance from the location of the reference profile. Each grid point within the average profile boxes is classified as signal if it is above 2% of the peak maximum. Finally, each profile is scaled such that the sum of its signal pixels normalizes to one. The analytical expression $\omega(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ defined in Section 11.3.2.3 for the expected intensity distribution is only a rough initial approximation, which is now replaced by the empirical reference profiles.

11.3.3.4. Intensity estimation

If an expected intensity distribution $\{p_i|i \in D_0\}$ of the observed profile is given in a domain D_0, the reflection intensity I can be estimated as

$$I=\frac{\sum_{i\in D}(c_i-b_i)p_i/v_i}{\sum_{i\in D}p_i^2/v_i},$$

which minimizes the function

$$\psi(I)=\sum_{i\in D}(c_i-I\cdot p_i-b_i)^2/v_i,\quad \sum_{i\in D_0}p_i=1.$$

b_i, c_i, v_i ($i \in D$) are the background, contents and variance of pixels observed in a subdomain $D \subseteq D_0$ of the expected distribution. The background b_i underneath a diffraction spot is often assumed to be a constant which is estimated from the neighbourhood around the reflection. Determination of reflection intensities by profile fitting has a long tradition (Diamond, 1969; Ford, 1974; Kabsch, 1988*b*; Otwinowski, 1993). Implementations of the method differ mainly in their assumptions about the variances v_i. Ford used constant variances, which work well for films, which have a high intrinsic background. In *XDS*, which was originally designed for a multwire detector, $v_i \propto p_i$ was assumed, which results in a straight summation of background-subtracted counts within the expected profile region, $I = \sum_{i\in D}(c_i - b_i)/\sum_{i\in D} p_i$. This particular simple formula is very satisfactory for the low background typical of these detectors. For the general case, however, better results can be obtained by using $v_i = b_i + Ip_i$ for the pixel variances as shown by Otwinowski and implemented in *DENZO* and in later versions of *XDS*. Starting with $v_i = b_i$, the intensity is now found by an iterative process which is terminated if the new intensity estimate becomes negative or does not change within a small tolerance, which is usually reached after three cycles. It can be shown that the solution thus obtained is unique.

11.3.4. Scaling

The integrated intensities of the reflections need to be corrected by various factors arising from the following:

(i) changes in the intensity of the incident beam and variations in the illuminated crystal volume,
(ii) absorption of incident and diffracted beams,
(iii) radiation damage,
(iv) variations in detector sensitivity within the detector plane, and
(v) different crystal sizes and crystalline order if the data are from several crystals.

The combined effect manifests itself in correlations of the intensity of a reflection with details of its measurement, such as time (or image number) and location in the detector plane. Usually, many statistically independent observations of symmetry-related reflections are recorded in the rotation images taken from one or several similar crystals of the same compound. The squared structure-factor amplitudes of equivalent reflections should be equal and many scaling procedures (see, for example Evans, 2006; Otwinowski *et al.*, 2003; Kabsch, 1988*b*) exploit this *a priori* knowledge to determine a correction factor for each observed intensity. However, the scaling programs differ in the details of their scaling models, *i.e.* the parameterization and methods used for determination of the correction factors. Below, the approach is described as implemented recently in the programs *XDS* and *XSCALE* (Chapter 11.6).

If more than one data set is included, these are first put on approximately the same scale by the factor $K{\cdot}\exp[B{\cdot}(2\sin\theta/\lambda^2)]$ involving two parameters, K and B, for each data set. The parameter values are assigned so that the resulting correction factors fit best to the observed intensity ratios of common reflections in each pair of data sets.

For the more detailed corrections, three types of two-dimensional functions are used in succession to remove correlations of the intensity of a reflection with (i) image number and resolution, (ii) location in the detector plane and (iii) image number and 13 detector surface regions. To correct for non-uniform detector response such as edge effects at the boundaries of multisegment detectors, the use of smooth analytical correction functions was avoided. Instead, the correction functions are sampled at a finite set of grid regions covering all of the function's definition range. The grid regions are chosen automatically to be as small as possible without overfitting the data so that each sampling region contains more than a specified minimum number of reflections (default 50). Thus, the correction function G is represented by a possibly large number of reciprocal factors G_l, where the subscript l denotes the grid regions.

The correction factors G_l are found in a cyclic prodedure starting with $G_l = 1$. In each cycle, G_l is updated by a factor g_l. The target function for refinement is based on an observational equation for each reflection,

$$\psi_{hl} = (I_{hl} - g_l G_l I_h)/\sigma_{hl},$$

as introduced by Hamilton *et al.* (1965). The subscript h represents the unique reflection indices and hl denotes symmetry-related reflections to h that need to be corrected by the reciprocal scaling factor g_l associated with grid point l; I_{hl} and σ_{hl} are their weighted mean and standard deviation, respectively. This standard deviation is considered to be infinitely large if no such reflection was measured, which amounts to omitting the observational equation altogether. The factors g_l and the 'true' intensities I_h are found at the minimum of the function

$$\Psi = \sum_{hl} \psi_{hl}^2 + \sum_l (g_l G_l - 1)^2/\sigma^2.$$

The first sum on the right side is a homogeneous function of g_l of degree zero so that the g_l would only be determined up to an arbitrary factor. The second sum on the right side is used to weakly restrain the scaling factors to one; a reasonable value is $\sigma = 0.05$. Minimization of Ψ leads to updates g_l in terms of the 'true' intensities I_h which again depend on g_l,

$$\begin{aligned} g_l &= \frac{\sum_h I_h (I_{hl}/G_l)/(\sigma_{hl}/G_l)^2 + G_l/\sigma^2}{\sum_h I_h^2/(\sigma_{hl}/G_l)^2 + G_l^2/\sigma^2} \\ I_h &= \frac{\sum_l g_l (I_{hl}/G_l)/(\sigma_{hl}/G_l)^2}{\sum_l g_l^2/(\sigma_{hl}/G_l)^2} \\ I_h^o &= \frac{\sum_l (I_{hl}/G_l)/(\sigma_{hl}/G_l)^2}{\sum_l 1/(\sigma_{hl}/G_l)^2}. \end{aligned}$$

The new update factors g_l are obtained by using 'true' intensities I_h^o from the previous cycle instead of the current I_h as defined above. At the end of the cycle, the old correction factors G_l are updated by multiplication with the new g_l. This cyclic procedure typically converges in less than six cycles.

The approach described here has been implemented in *XDS* and *XSCALE* and has been successfully used for more than two years. In contrast to the 'shortest path' eigenvector method of Fox & Holmes (1966), which is very efficient for a relatively small number of variables, the computations here require a time that is proportional to the number of reflections used for scaling and thus quickly lead to a solution even when a very large number of correction factors from many data sets are involved.

A recent literature search revealed that the above method of alternating optimization steps for the intensities and scaling factors is essentially the method of Monahan *et al.* (1967). Their original work was largely neglected, presumably because of the lack of convergence of the iterations in some test cases. Apparently, the preliminary scaling step in *XSCALE* is essential for convergence. A detailed mathematical analysis of the alternate convex search method has been carried out (Junk *et al.*, 2011).

11.3.5. Post refinement

The number of fully recorded reflections on each single image rapidly declines for small oscillation ranges and the complete intensities of the partially recorded reflections have to be estimated. This presented a serious obstacle in early structural work on virus crystals, as the crystal had to be replaced after each exposure on account of radiation damage. A solution to this problem, the 'post-refinement' technique, was found by Schutt, Winkler and Harrison and variants of this powerful method have been incorporated into most data-reduction programs (for a detailed discussion, see Harrison *et al.*, 1985; Rossmann, 1985). The method derives complete intensities of reflections that are only partially recorded on an image from accurate estimates for the fractions of observed intensity: the 'partiality'. The partiality of each reflection can always be calculated as a function of orientation, unit-cell metric, mosaic spread of the crystal and model intensity distributions. The accuracy of the estimated full reflection intensity obviously then strongly depends on a precise knowledge of the parameters describing the diffraction experiment. Usually, symmetry-related fully recorded reflections can be found for many of the partial reflections and the list of such pairs of intensity observations can be used to refine the required parameters using a least-squares procedure. Clearly, this refinement is carried out after all images have been processed, which explains why the procedure is called 'post refinement'.

Adjustments of the diffraction parameters s_μ ($\mu = 1, \ldots, k$) are determined by minimization of the function E, which is defined as the weighted sum of squared residuals between calculated and observed partial intensites.

$$\begin{aligned} E &= \sum_{hj} w_{hj}(\Delta_{hj})^2 \\ \Delta_{hj} &= R_j(\varphi_{hj})g_j I_h - I_{hj} \\ w_{hj} &= 1/\{\sigma^2(I_{hj}) + [R_j(\varphi_{hj})g_j]^2\sigma^2(I_h)\}. \end{aligned}$$

Here, I_{hj} is the intensity recorded on image j of a partial reflection with indices summarized as hj, I_h is the mean of the observed intensities of all fully recorded reflections symmetry-equivalent to hj, g_j is the inverse scaling factor of image j, φ_{hj} is the calculated spindle angle of reflection hj at diffraction and R_j is the computed fraction of total intensity recorded on image j.

Expansion of the residuals Δ_{hj} to first order in the parameter changes δs_μ and minimization of $E(\delta s_\mu)$ leads to the k normal equations

$$\sum_{\mu'=1}^{k}\left(\sum_{hj} w_{hj}\frac{\partial\Delta_{hj}}{\partial s_\mu}\frac{\partial\Delta_{hj}}{\partial s_{\mu'}}\right)\delta s_{\mu'} = -\sum_{hj} w_{hj}\Delta_{hj}\frac{\partial\Delta_{hj}}{\partial s_\mu}.$$

Often, the normal matrix is ill-conditioned since changes in some unit-cell parameters or small rotations of the crystal about the incident X-ray beam do not significantly affect the calculated partiality R_j. To take care of these difficulties, the system of equations is rescaled to yield unit diagonal elements for the normal matrix and the correction vector δs_μ is filtered by projection into a subspace defined by the eigenvectors of the normal matrix with sufficiently large eigenvalues (Diamond, 1966).

The parameters are corrected by the filtered δs_μ and a new cycle of refinement is started until a minimum of E is reached. The weights, residuals and their gradients are calculated using the current values for s_μ and g_j at the beginning of each cycle. The derivatives

$$\frac{\partial\Delta_{hj}}{\partial s_\mu} = g_j I_h\left(\frac{\partial R_j}{\partial\varphi_{hj}}\frac{\partial\varphi_{hj}}{\partial s_\mu} + \frac{\partial R_j}{\partial\sigma_{\rm M}}\frac{\partial\sigma_{\rm M}}{\partial s_\mu} + \frac{\partial R_j}{\partial|\zeta_{hj}|}\frac{\partial|\zeta_{hj}|}{\partial s_\mu}\right)$$

appearing in the normal equations can be worked out from the definitions given in Sections 11.3.2.2 and 11.3.2.4 (to simplify the following equations, the subscript hj is omitted). The fraction R_j of total intensity can be expressed in terms of the error function (see Section 11.3.2.4) as

$$\begin{aligned} R_j &= [\mathrm{erf}(z_1) - \mathrm{erf}(z_2)]/2 \\ z_1 &= |\zeta|(\varphi_0 + j\Delta_\varphi - \varphi)/2^{1/2}\sigma_M \\ z_2 &= |\zeta|[\varphi_0 + (j-1)\Delta_\varphi - \varphi]/2^{1/2}\sigma_M. \end{aligned}$$

Using the relation $\mathrm{d}\,\mathrm{erf}(z)/\mathrm{d}z = (2/\pi^{1/2})\exp(-z^2)$, the derivatives of R_j are

$$\begin{aligned} \frac{\partial R_j}{\partial\varphi} &= [\exp(-z_2^2) - \exp(-z_1^2)]|\zeta|/[\sigma_M(2\pi)^{1/2}] \\ \frac{\partial R_j}{\partial\sigma_M} &= [z_2\exp(-z_2^2) - z_1\exp(-z_1^2)]/(\sigma_M\pi^{1/2}) \\ \frac{\partial R_j}{\partial|\zeta|} &= [z_1\exp(-z_1^2) - z_2\exp(-z_2^2)]/(|\zeta|\pi^{1/2}). \end{aligned}$$

The derivatives $\partial\varphi/\partial s_\mu$, $\partial\sigma_M/\partial s_\mu$ and $\partial|\zeta|/\partial s_\mu$ remain to be worked out (not shown here). As discussed in detail by Greenhough & Helliwell (1982), spectral dispersion and asymmetric beam crossfire lead to some variation in σ_M, which makes it necessary to include additional parameters in the list s_μ. The effect of these parameters on the partiality is dealt with easily by the derivatives $\partial\sigma_M/\partial s_\mu$.

The refinement scheme described above requires initial scaling factors g_j. With the now improved estimates for the partialities R_j, a new set of scaling factors can be obtained using the method outlined in Section 11.3.4. This alternating procedure of scaling and post refinement usually converges within three cycles.

The use of error functions for modelling partiality, as implicated by a Gaussian model for describing spot shape, was chosen here for reasons of conceptual simplicity and coherence. This choice is unlikely to significantly alter the results of post refinement that are based on other functions of similar form (see the discussion by Rossmann, 1985).

11.3.6. Space-group assignment

Identification of the correct space group is not always an easy task and should be postponed for as long as possible. Sometimes, the true space group only becomes known when the structure has been successfully solved and refined! However, one can expect to identify a small number of possibilities from the diffraction experiment.

Fortunately, all data processing as implemented in the program *XDS* can be carried out in the absence of any knowledge of the crystal symmetry and unit-cell parameters. In this case, a reduced cell is extracted from the observed diffraction pattern and processing of the data images continues to completion as if the crystal were triclinic. Clearly, the reflection indices then refer to the reduced cell and must be reindexed once the space group is known. For all space groups, the required reindexing transformation is linear and involves only whole numbers, as shown in Part 9 of Volume A of *International Tables for Crystallography* (2005).

Automatic space-group assignment is carried out in two steps once integrated intensities of all reflections are available (see Chapter 11.6). Firstly, the Bravais lattices are identified that are compatible with the reduced cell derived from the observed diffraction pattern. In the second step, all enantiomorphous space groups compatible with the observed lattice symmetry are rated by a redundancy-independent R factor. The group is selected that explains all integrated intensities in the data set at an acceptable R factor requiring a minimum number of unique reflections (Occam's principle). This approach deliberately avoids any test for the presence of screw axes as these tests would depend strongly on the completeness of the data. Fortunately, the presence or absence of screw axes is irrelevant for the determination of data correction/scaling factors (see Section 11.3.4).

Table 11.3.6.1

Rating of lattice types implied by a given reduced cell

Lattice type		Quality index	Conventional unit-cell parameters (Å, °)						Reindexing transformation
			a	b	c	α	β	γ	
44	*aP*	0.0	159.3	159.4	160.4	90.1	90.1	90.1	$11\bar{1}0/1\bar{1}10/\bar{1}\bar{1}\bar{1}0$
31	*aP*	0.4	159.3	159.4	160.4	90.1	89.9	89.9	1000/0100/0010
34	*mP*	1.4	159.3	160.4	159.4	90.1	90.1	90.1	$\bar{1}000/00\bar{1}0/0\bar{1}00$
14	*mC*	1.4	225.1	225.6	160.4	90.0	90.1	89.9	$1100/\bar{1}100/0010$
33	*mP*	1.5	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
35	*mP*	2.0	159.4	159.3	160.4	90.1	90.1	90.1	$0\bar{1}00/\bar{1}000/00\bar{1}0$
13	*oC*	2.3	225.1	225.6	160.4	90.0	90.1	89.9	$1100/\bar{1}100/0010$
32	*oP*	2.4	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
10	*mC*	2.5	225.1	225.6	160.4	90.0	90.1	90.1	$\bar{1}\bar{1}00/\bar{1}100/00\bar{1}0$
11	***tP***	**3.4**	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
25	*mC*	5.9	226.0	226.2	159.3	90.0	90.2	89.7	$0110/0\bar{1}10/1000$
20	*mC*	6.4	226.0	226.2	159.3	90.0	90.2	90.3	$0\bar{1}\bar{1}0/0\bar{1}10/\bar{1}000$
4	*hR*	7.4	225.6	226.2	276.2	90.3	89.9	119.9	$1\bar{1}00/\bar{1}010/\bar{1}\bar{1}\bar{1}0$
23	*oC*	7.8	226.0	226.2	159.3	90.0	90.2	89.7	$0110/0\bar{1}10/1000$
3	*cP*	7.8	159.3	159.4	160.4	90.1	90.1	90.1	1000/0100/0010
21	*tP*	8.2	159.4	160.4	159.3	90.1	90.1	90.1	0100/0010/1000
2	*hR*	8.7	225.1	225.8	276.9	90.2	90.0	119.8	$1100/\bar{1}0\bar{1}0/\bar{1}110$
5	*cI*	173.6	225.8	225.1	226.0	60.2	59.9	60.2	1010/1100/0110

11.3.6.1. Determination of the Bravais lattice

The determination of possible Bravais lattices is based upon the concept of the reduced cell whose metric parameters characterize 44 lattice types as described in Part 9 of Volume A of *International Tables for Crystallography* (2005). A primitive basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ of a given lattice is defined there as a reduced cell if it is right-handed and if the components of its metric tensor

$$A = \mathbf{b}_1 \cdot \mathbf{b}_1, \quad B = \mathbf{b}_2 \cdot \mathbf{b}_2, \quad C = \mathbf{b}_3 \cdot \mathbf{b}_3$$
$$D = \mathbf{b}_2 \cdot \mathbf{b}_3, \quad E = \mathbf{b}_1 \cdot \mathbf{b}_3, \quad F = \mathbf{b}_1 \cdot \mathbf{b}_2$$

satisfy a number of conditions (inequalities). The main conditions state that the basis vectors are the shortest three linear independent lattice vectors with either all acute or all non-acute angles between them. As specified in *International Tables for Crystallography*, each of the 44 lattice types is characterized by additional equality relations among the six components of the reduced-cell metric tensor. As an example, for lattice character 11 (Bravais type *tP*) the components of the metric tensor of the reduced cell must satisfy

$$A = B, \quad B \leq C, \quad D = 0, \quad E = 0, \quad F = 0.$$

(Note that the other tetragonal primitive lattice character 21 requires $A \leq B = C$ with the fourfold as the shortest axis.)

Any primitive triclinic cell describing a given lattice can be converted into a reduced cell. It is well known, however, that the reduced cell thus derived is sensitive to experimental error. Hence, the direct approach of first deriving the correct reduced cell and then finding the lattice type is unstable and may in certain cases even prevent identification of the correct Bravais lattice.

A suitable solution of the problem has been found that avoids any decision as to what the 'true' reduced cell is (see Kabsch, 1993). The essential ingredients of this procedure are (i) a database of possible reduced cells and (ii) a backward-search strategy that finds the best-fitting cell in the database for each lattice type.

The database is derived from a seed cell which strictly satisfies the definitions for a reduced cell. All cells of the same volume as the seed cell whose basis vectors can be linearly expressed in terms of the seed vectors by indices −1, 0 or +1 are included in the database. Each unit cell in the database is considered as a potential reduced cell, although some of the defining conditions as given in Part 9 of Volume A of *International Tables for Crystallography* (2005) may be violated. These violations are treated as arising from experimental error.

The backward-search strategy starts with the hypothesis that the lattice type is already known and identifies the best-fitting cell in the database of possible reduced cells. In contrast to a forward-directed search, it is now always possible to decide which conditions have to be satisfied by the components of the metric tensor of the reduced cell. The total amount by which all these equality and inequality conditions are violated is used as a quality index. For example, to find out how well a potential reduced cell $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ from the database characterizes lattice character 11 (Bravais lattice *tP*), the quality index

$$p_{11}(\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3) = |A - B| + \max(0, B - C) + |D| + |E| + |F|$$

is computed. Positive values of p_{11} indicate that some conditions are not satisfied. All potential reduced cells in the database are tested and the smallest value for p_{11} is used for rating lattice type 11. A similar test is carried out for all 44 possible lattice types using quality indices based on their defining conditions as listed in Part 9 of Volume *A* of *International Tables for Crystallography* (2005).

The results obtained using this method are shown in Table 11.3.6.1 for the example of a data set comprising 177 images with each exposure covering 0.5° of spindle rotation. The space group of the protein crystal was $P4_32_12$ (unit-cell parameters a = 159.4, b = 159.4, c = 160.3 Å), but this knowledge was not used in the processing. Instead, the data were processed with respect to a triclinic reduced cell derived from the observed diffraction pattern as described above. The images contained a total of 292 998 reflections within the resolution range 20.0–3.0 Å; 57 548 reflections in the resolution range 10.0–5.0 Å were used for space-group determination. For determination of the lattice symmetry all 44 possibilities were considered and rated by their quality index. The table shows the possible lattice symmetries, their implied conventional unit-cell parameters and a reindexing transformation. The table entries are sorted by increasing quality index and reveal a nearly cubic lattice symmetry. A lattice symmetry is considered to be acceptable if it has a low quality

Table 11.3.6.2
Identification of possible space groups

Space group		Lattice type		R_{meas} (%)	UNIQUE	COMPARED	Conventional unit-cell parameters (Å, °)					
							a	b	c	α	β	γ
1	$P1$	44	aP	5.8	35341	22207	159.3	159.4	160.4	90.1	90.1	90.1
1	$P1$	31	aP	5.8	35341	22207	159.3	159.4	160.4	90.1	89.9	89.9
3	$P2$	33	mP	6.5	21904	35644	159.3	159.4	160.4	90.0	90.1	90.0
3	$P2$	34	mP	7.0	26743	30805	159.3	160.4	159.4	90.0	90.1	90.0
5	$C2$	10	mC	7.7	22207	35341	225.1	225.6	160.4	90.0	90.1	90.0
5	$C2$	14	mC	7.7	22207	35341	225.1	225.6	160.4	90.0	90.1	90.0
16	$P222$	32	oP	7.9	14461	43087	159.3	159.4	160.4	90.0	90.0	90.0
21	$C222$	13	oC	8.0	15094	42454	225.1	225.6	160.4	90.0	90.0	90.0
3	$P2$	35	mP	8.2	25786	31762	159.4	159.3	160.4	90.0	90.0	90.0
75	$P4$	11	tP	8.5	14944	42604	159.4	159.4	160.4	90.0	90.0	90.0
89	$\boldsymbol{P422}$	**11**	$\boldsymbol{tP}$	**9.0**	8086	49462	159.4	159.4	160.4	90.0	90.0	90.0
146	$R3$	2	hR	45.2	20068	37480	225.5	225.5	276.9	90.0	90.0	120.0
5	$C2$	20	mC	46.9	23125	34423	226.0	226.2	159.3	90.0	90.2	90.0
5	$C2$	25	mC	46.9	23125	34423	226.0	226.2	159.3	90.0	90.2	90.0
75	$P4$	21	tP	49.2	14828	42720	159.9	159.9	159.3	90.0	90.0	90.0
89	$P422$	21	tP	50.7	7876	49672	159.9	159.9	159.3	90.0	90.0	90.0
21	$C222$	23	oC	51.3	15155	42393	226.0	226.2	159.3	90.0	90.0	90.0
195	$P23$	3	cP	57.3	5344	52204	159.7	159.7	159.7	90.0	90.0	90.0
207	$P432$	3	cP	58.1	2896	54652	159.7	159.7	159.7	90.0	90.0	90.0
155	$R32$	4	hR	59.7	9038	48510	225.9	225.9	276.2	90.0	90.0	120.0
155	$R32$	2	hR	60.7	10487	47061	225.5	225.5	276.9	90.0	90.0	120.0
146	$R3$	4	hR	61.1	16751	40797	225.9	225.9	276.2	90.0	90.0	120.0

index and its implied unit-cell parameters do not violate the ideal values by more than 3.0° in angles and 3% in cell axes. Thus, except for the last entry, all of the lattice symmetries in the table are acceptable; the correct lattice type 11 *tP* is highlighted. Lattice symmetries that are not accepted include all body-centred lattices or those that are centred on all faces; they are omitted from the table.

The reindexing transformation REIDX() consists of 12 integers that relate the original indices h, k, l used during the integration to the indices h', k', l' with respect to the new cell.

$$\begin{aligned} h' &= \frac{[\text{REIDX}(1)\times h + \text{REIDX}(2)\times k + \text{REIDX}(3)\times l]}{\text{IDXV}} \\ &\quad + \text{REIDX}(4) \\ k' &= \frac{[\text{REIDX}(5)\times h + \text{REIDX}(6)\times k + \text{REIDX}(7)\times l]}{\text{IDXV}} \\ &\quad + \text{REIDX}(8) \\ l' &= \frac{[\text{REIDX}(9)\times h + \text{REIDX}(10)\times k + \text{REIDX}(11)\times l]}{\text{IDXV}} \\ &\quad + \text{REIDX}(12). \end{aligned}$$

The value of the integer IDXV depends on the lattice type used for specifying reflection indices in the integration step. IDXV is 1 for a primitive lattice, 2 for a face-centred or body-centred lattice, 3 for a rhombohedral lattice and 4 for a lattice centred on all faces. In the example case we have IDXV = 1 because integration was carried out in space group $P1$.

Note also that elements 4, 8 and 12 of the transformation are always 0 in this example. These three extra elements were introduced to provide a simple tool for correcting the indices if all reflections are misindexed by a constant.

11.3.6.2. Finding possible space groups

For protein crystals, the absence of parity-changing symmetry operators restricts the number of possible space groups to 65 instead of 230. Moreover, the determination of correction factors for the integrated intensities does not depend on the presence or absence of any screw axes so that data processing can be finished without this knowledge. This reduces the problem to the identification of an enantiomorphous space group without screw axes that is compatible with the observed lattice symmetry (see above).

For solution of the problem, a quality indicator of the mean variation in the intensities of symmetry-equivalent reflections (R_{meas}) is calculated for each possible group. The decision for a particular group is then based on Occam's principle: the selected group must explain all integrated intensities in the data set at acceptable quality, thereby requiring a minimum number of unique reflections.

A suitable redundancy-independent data quality indicator has been suggested by Diederichs & Karplus (1997) and Weiss (2001),

$$R_{\text{meas}} \equiv R_{\text{r.i.m.}} = \frac{\sum_{hl}\left[n_h/(n_h-1)\right]^{1/2}|I_{hl}-I_h|}{\sum_{hl} I_{hl}}.$$

The subscript h represents the unique reflection indices and hl denotes any of the n_h symmetry-related reflections to h. The absolute differences between the observed intensities I_{hl} and their mean intensity I_h are weighted to remove any dependency on n_h and compared with the intensities. Small values of R_{meas} indicate accurate single observations I_{hl} and the use of symmetry operators compatible with the intensity data set.

For the above example data set, Table 11.3.6.2 lists all enantiomorphous groups which are in harmony with the observed lattice symmetry shown in Table 11.3.6.1. For each listed space group, UNIQUE is the number of unique reflections and COMPARED is the number of reflections used to calculate the redundancy-independent R factor R_{meas}. Two sets of groups can be distinguished clearly: those implying an acceptable R_{meas} and a second set with $R_{\text{meas}} > 45\%$, which is totally unacceptable. Among the acceptable solutions a minimum number of unique reflections is needed if the crystal has the tetragonal space-group symmetry $P422$.

This chapter is based on a recent article published in *Acta Crystallographica Section D* (Kabsch, 2010).

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions.* New York: Dover Publications.

Bricogne, G. (1986). *Indexing and the Fourier transform*. In *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software (Phase III)*, p. 28. Paris: LURE.

Diamond, R. (1966). *A mathematical model-building procedure for proteins*. *Acta Cryst.* **21**, 253–266.

Diamond, R. (1969). *Profile analysis in single crystal diffractometry*. *Acta Cryst.* A**25**, 43–55.

Diederichs, K. & Karplus, P. A. (1997). *Improved R-factors for diffraction data analysis in macromolecular crystallography Nat. Struct. Biol.* **4**, 269–275.

Dijkstra, E. W. (1976). *A Discipline of Programming*, pp. 154–167. New Jersey: Prentice–Hall.

Evans, P. (2006). *Scaling and assessment of data quality*. *Acta Cryst.* D**62**, 72–82.

Ford, G. C. (1974). *Intensity determination by profile fitting applied to precession photographs*. *J. Appl. Cryst.* **7**, 555–564.

Fox, G. C. & Holmes, K. C. (1966). *An alternative method of solving the layer scaling equations of Hamilton, Rollett and Sparks*. *Acta Cryst.* **20**, 886–891.

Greenhough, T. J. & Helliwell, J. R. (1982). *Oscillation camera data processing: reflecting range and prediction of partiality. I. Conventional X-ray sources*. *J. Appl. Cryst.* **15**, 338–351.

Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *On the relative scaling of X-ray photographs*. *Acta Cryst.* **18**, 129–130.

Harrison, S. C., Winkler, F. K., Schutt, C. E. & Durbin, R. M. (1985). *Oscillation method with large unit cells*. *Methods Enzymol.* **114**, 211–237.

Howard, A. (1986). *Autoindexing*. In *Proceedings of the EEC Cooperative Workshop on Position-Sensitive Detector Software (Phases I and II)*, pp. 89–94 Paris: LURE.

International Tables for Crystallography (2005). Volume A, *Space-Group Symmetry*, pp. 742–760. Heidelberg: Springer.

Junk, M., Diederichs, K. & Kabsch, W. (2011). *Analysis of the alternate convex search method for scaling crystallographic data. Comput. Optim. Appl*. Submitted.

Kabsch, W. (1988*a*). *Automatic indexing of rotation diffraction patterns*. *J. Appl. Cryst.* **21**, 67–71.

Kabsch, W. (1988*b*). *Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector*. *J. Appl. Cryst.* **21**, 916–924.

Kabsch, W. (1993). *Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants*. *J. Appl. Cryst.* **26**, 795–800.

Kabsch, W. (2010). *Integration, scaling, space-group assignment and post-refinement*. *Acta Cryst.* D**66**, 133–144.

Monahan, J. E., Schiffer, M. & Schiffer, J. P. (1967). *On the scaling of X-ray photographs*. *Acta Cryst.* **22**, 322.

Otwinowski, Z. (1993). *Oscillation data reduction program*. In *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.

Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Multiparametric scaling of diffraction intensities*. *Acta Cryst.* A**59**, 228–234.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode*. *Methods Enzymol.* **276**, 307–326.

Pflugrath, J. W. (1997). *Diffraction-data processing for electronic detectors: theory and practice*. *Methods Enzymol.* **276**, 286–306.

Rossmann, M. G. (1985). *Determining the intensity of Bragg reflections from oscillation photographs*. *Methods Enzymol.* **114**, 237–280.

Schutt, C. & Winkler, F. K. (1977). *The oscillation method for very large unit cells*. In *The Rotation Method in Crystallography*, edited by U. W. Arndt & A. J. Wonacott, pp. 173–186. Amsterdam: North-Holland.

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *An algorithm for automatic indexing of oscillation images using Fourier analysis*. *J. Appl. Cryst.* **30**, 1036–1040.

Weiss, M. S. (2001). *Global indicators of X-ray data quality*. *J. Appl. Cryst.* **34**, 130–135.

Wirth, N. (1976). *Algorithms + Data Structures = Programs*, pp. 264–274. New York: Prentice–Hall.

International Tables for Crystallography (2012). Vol. F, Chapter 11.4, pp. 282–295.

Chapter 11.4. *DENZO* and *SCALEPACK*

Z. Otwinowski, W. Minor, D. Borek and M. Cymborowski

11.4.1. Introduction

X-ray diffraction data analysis, as performed by the *HKL* package (Otwinowski, 1993; Otwinowski & Minor, 1997) or similar programs (Rossmann, 1979; Leslie & Tsukihara, 1980; Howard *et al.*, 1985; Blum *et al.*, 1987; Bricogne, 1987; Howard *et al.*, 1987; Messerschmidt & Pflugrath, 1987; Kabsch, 1988; Higashi, 1990; Sakabe, 1991), is used to obtain the following results:

(1) estimates of structure-factor amplitudes and determination of the crystal symmetry;
(2) estimates of the crystal unit-cell parameters;
(3) estimates of the errors of structure-factor amplitudes and unit-cell parameters;
(4) detector calibration; and
(5) detection of hardware malfunctions.

Other results, such as the indexing of diffraction patterns, are in most cases only intermediate steps in the process of reaching the above goals. The *HKL* system and other programs have tools to validate these intermediate results by self-consistency checks.

The fundamental stages of data analysis are:

(1) inspection of the diffraction images;
(2) (auto)indexing;
(3) refinement of parameters describing diffraction geometry;
(4) integration of the diffraction peaks;
(5) conversion of the diffraction-peak intensities to a common scale;
(6) determination of the diffraction-pattern symmetry and merging of symmetry-related reflections; and
(7) assessment of structural and phasing signals, summary statistics and estimation of errors.

The order of these stages represents the natural flow of data reduction, but quite often these steps are repeated iteratively to introduce information gained at a later stage.

The three basic questions in collecting diffraction data are:

(1) *whether to collect data from a particular sample*;
(2) *what range of data to collect*; and
(3) *how to collect and analyse the data.*

These questions and steps (1)–(7) of data analysis are intimately intertwined.

Data analysis makes specific assumptions which the collected data should satisfy. However, the experimenter can verify whether the data satisfy those assumptions only by data analysis. This circular logic can be broken by an iterative process. Data analysis in real time provides immediate feedback during data collection and can remove the guesswork about *whether*, *what* and *how* from the process. The description of data analysis and algorithms that follows will make frequent references to the assumptions about the data and offer guidelines on how to make the experiment fulfil these assumptions.

This article uses the *HKL* package coordinate system to describe data algorithms and analysis. However, as most equations are written in vector notation, they can be easily adapted to conventions used in other programs.

11.4.2. Diffraction from a perfect crystal lattice

X-ray photons can scatter from individual electrons by inelastic and incoherent processes. The coherent scattering by the whole crystal is called diffraction.[1] Energy conservation, when expressed in photon momentum vectors, is equivalent to

$$\mathbf{S}\cdot\mathbf{S}_0 = \tfrac{1}{2}\mathbf{S}\cdot\mathbf{S}, \qquad (11.4.2.1)$$

where $\mathbf{S}$ is the diffraction vector, defined as the change of photon momentum in the scattering process, and $\mathbf{S}_0$ is the incident-beam vector, which is oriented in the direction of the beam and has length $1/\lambda$. Diffraction from a perfect crystal lattice occurs when scattering from all repeating crystal elements is in phase, which can be stated in vector algebra as

$$\mathbf{S}\cdot\mathbf{a} = h, \qquad (11.4.2.2)$$

$$\mathbf{S}\cdot\mathbf{b} = k, \qquad (11.4.2.3)$$

$$\mathbf{S}\cdot\mathbf{c} = l, \qquad (11.4.2.4)$$

where h, k, l are integers called Miller indices and $\mathbf{a}, \mathbf{b}, \mathbf{c}$ are the real-space crystal periodicity vectors, defining the orientation matrix:

$$[\mathbf{A}] = \begin{pmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{pmatrix}.$$

In shorter notation, equations (11.4.2.2)–(11.4.2.4) may be written as $\mathbf{h} = [\mathbf{A}]\mathbf{S}$, which in reciprocal space is equivalent to $\mathbf{S} = [\mathbf{A}]^{-1}\mathbf{h}$, with the orientation matrix defined as the inverse of $[\mathbf{A}]$.

The condition for crystal diffraction with Miller indices h, k, l is the existence of a (unique) vector $\mathbf{S}$ which is a solution to equations (11.4.2.1)–(11.4.2.4). Equation (11.4.2.1) states the diffraction condition for vector $\mathbf{S}$. The space of the solutions to equations (11.4.2.2)–(11.4.2.4) is called reciprocal space, and vector $\mathbf{S}$ belongs to this space. However, the following presentation does not depend on the properties of reciprocal space. The laboratory coordinate system used here has its origin at the position of the crystal. A diffraction peak at the detector position in three-dimensional laboratory space $\mathbf{X} = \{x, y, z\}$ corresponds to vector $\mathbf{S}$:

$$\mathbf{S} = \mathbf{X}/\lambda|\mathbf{X}| + \mathbf{S}_0. \qquad (11.4.2.5)$$

Rotation of the crystal around the goniostat axes can be described by vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$ [equations (11.4.2.2)–(11.4.2.4)] as a function of the goniostat angles $\varphi_1, \varphi_2, \varphi_3$:

$$\mathbf{a} = [\mathbf{R}_1(\varphi_1)][\mathbf{R}_2(\varphi_2)][\mathbf{R}_3(\varphi_3)]\mathbf{a}_0, \qquad (11.4.2.6)$$

where the vectors $\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0$ represent the crystal orientation at the zero position of the goniostat. These rotations can be described as (Bricogne, 1987)

---

[1] Owing to the large difference in mass between the crystal and the photon, the energy of the photon is virtually unchanged.

Copyright © 2012 International Union of Crystallography

$$[\mathbf{R}(\varphi)] = \cos(\varphi)\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + [1-\cos(\varphi)]\begin{pmatrix} e_x e_x & e_x e_y & e_x e_z \\ e_x e_y & e_y e_y & e_y e_z \\ e_x e_z & e_z e_y & e_z e_z \end{pmatrix}$$
$$+ \sin(\varphi)\begin{pmatrix} 0 & -e_z & e_y \\ e_z & 0 & -e_x \\ -e_y & e_x & 0 \end{pmatrix}, \qquad (11.4.2.7)$$

where the values e_x, e_y, e_z represent the direction cosines of a rotation axis. To complete the description of the diffraction geometry, we need a function $\mathbf{X}(p, q)$, describing the position in experimental space of each pixel of the detector with integer coordinates $\{p, q\}$. This function is detector-specific and includes the description of the detector geometry and distortion. For a planar detector

$$\mathbf{X}(p, q) = [\mathbf{R}_x][\mathbf{R}_y][\mathbf{R}_z]\big([\mathbf{R}_{2\theta}]([\mathbf{L}](\mathbf{K}[\mathbf{D}(p, q)] - \mathbf{B}) + \mathbf{T}_D) - \mathbf{T}_D\big)$$
$$+ \mathbf{T}_D, \qquad (11.4.2.8)$$

where $\mathbf{R}_x$, $\mathbf{R}_y$, $\mathbf{R}_z$ are matrices describing the detector's departure from a nominal orientation, $\mathbf{R}_{2\theta}$ is the rotation around the 2θ axis, $\mathbf{T}_D$ is the translation between the detector and the crystal, $\mathbf{L}$ is an operator describing one of the eight possible axis-direction conventions used by the detector manufacturer, $\mathbf{K}$ is an operator scaling pixels to millimetres, $\mathbf{D}$ is a detector-specific distortion function, and $\mathbf{B}$ is the beam position on the detector surface.

Equations (11.4.2.1)–(11.4.2.8) define the necessary and sufficient conditions needed to fully describe both the existence and position on the detector of the diffraction peaks, and provide all information needed to perform autoindexing.

11.4.3. Autoindexing

Among the autoindexing algorithms that have been proposed (Vriend & Rossmann, 1987; Kabsch, 1988; Kim, 1989; Higashi, 1990; Leslie, 1993), the method based on periodicity of the reciprocal lattice tends to be the most reliable (Otwinowski & Minor, 1997; Steller *et al.*, 1997).

Autoindexing starts with a peak search, which results in a set of $\{p, q, i\}$ triplets, where i is the number of the image in which the peak with position $\{p, q\}$ was found. For any given rotation matrix $\mathbf{R}$

$$([\mathbf{R}]\mathbf{S}) \cdot ([\mathbf{R}]\mathbf{a}) = \mathbf{S} \cdot \mathbf{a}, \qquad (11.4.3.1)$$

where $\mathbf{R}$ is defined as

$$[\mathbf{R}] = \big[\mathbf{R}_3(-\varphi_3)\big]\big[\mathbf{R}_2(-\varphi_2)\big]\big[\mathbf{R}_1(-\varphi_1)\big]. \qquad (11.4.3.2)$$

When equation (11.4.3.1) is applied to equation (11.4.2.2), it results in

$$([\mathbf{R}]\mathbf{S}) \cdot \mathbf{a}_o = h, \qquad (11.4.3.3)$$

where $\mathbf{a}_o$ is a three-dimensional vector with as yet unknown components. Note that the matrix $[\mathbf{R}]$ represents crystal rotation when the crystal is in the diffraction condition defined by the existence of the solution to equations (11.4.2.1)–(11.4.2.4), described by vector $\mathbf{S}$. For data collected in wide oscillation mode,[2] the angle at which diffraction occurs is not known *a priori*; however, it can be approximated by the middle of the oscillation range of the image. Combining the peak position $\{p, q\}$ with equations (11.4.2.5) and (11.4.2.8) provides an estimate of the vector $\mathbf{S}$. Therefore, equation (11.4.3.3) and similar equations for the k and l Miller indices are approximately satisfied, given approximation and experimental errors. The purpose of autoindexing is to determine the unknown vectors $\mathbf{a}_0$, $\mathbf{b}_0$, $\mathbf{c}_0$ and the $\{h, k, l\}$ triplet for each diffraction peak. To accomplish this, three equations for each peak [equation (11.4.3.3) and the analogous equations for k and l] must be solved. *DENZO* introduced a method based on the observation that the maxima of the function

$$\sum_i \cos\big[2\pi([\mathbf{R}]\mathbf{S}_i) \cdot \mathbf{a}_0\big] \qquad (11.4.3.4)$$

are the approximate solutions to this set of equations. To speed up the search for all significant maxima, a two-step process is used. The first step is the search for maxima of function (11.4.3.4) on a three-dimensional uniform grid, made very fast owing to the use of a fast Fourier transform (FFT) to evaluate function (11.4.3.4). Function (11.4.3.4) is identical to structure-factor calculations in the space group $P1$, which allows the use of the crystallographic FFT. Because the maxima at the grid points (*HKL* uses a 128 × 128 × 128 grid) only approximate the maxima of function (11.4.3.4), the vectors resulting from a grid search are optimized by Newton's method. Function (11.4.3.4) has maxima not only for basic periodic vectors $\mathbf{a}_0$, $\mathbf{b}_0$ and $\mathbf{c}_0$, but also for any integer linear combination of them. Any set of three such vectors with a minimal nonzero determinant can be used to describe the crystal lattice. Steller *et al.* (1997) describe the algorithm that finds the most reliable set of three vectors. This set needs to be converted to the one conventionally used by crystallographers, as defined in *International Tables for Crystallography* Volume A (*IT* A) (2005).

To generate the conventional solution, a two-step procedure is used. In step 1, the reduced primitive triclinic cell is found using the algorithm provided by *IT* A. Subsequently, step 2 finds conventional cells in the other Bravais lattices of higher symmetry.

11.4.3.1. Lattice symmetry

The relationship between a higher-symmetry cell and the reduced primitive triclinic cell can be described by

$$[\mathbf{A}] = [\mathbf{M}][\mathbf{P}], \qquad (11.4.3.5)$$

where $[\mathbf{A}]$ and $[\mathbf{P}]$ are matrices of the type $\{\mathbf{a}_0, \mathbf{b}_0, \mathbf{c}_0\}$, with $[\mathbf{P}]$ representing the reduced triclinic primitive cell, and $[\mathbf{M}]$ is one of the 44 matrices listed in *IT* A. If $[\mathbf{A}]$ is generated using equation (11.4.3.5) from an experimentally determined $[\mathbf{P}]$, owing to experimental errors it will not exactly satisfy the symmetry restraints. *DENZO* introduced a novel index to evaluate the significance of this violation of symmetry. This index is based on the observation that from $[\mathbf{A}]$ one can deduce the parameters of the unit cell, apply symmetry restraints to the unit cell and calculate a matrix $[\mathbf{A}']$ for the unit cell that satisfies these symmetry restraints. If $[\mathbf{A}]$ perfectly satisfies the symmetry restraints, the matrix $[\mathbf{U}]$, where

$$[\mathbf{U}] = [\mathbf{A}][\mathbf{A}']^{-1}, \qquad (11.4.3.6)$$

will be unitary and

$$[\mathbf{U}]^T - [\mathbf{U}]^{-1} = 0. \qquad (11.4.3.7)$$

The index of distortion calculated and presented by *DENZO* is

$$\left\{\sum_i \sum_j \left([\mathbf{U}]^T_{ij} - [\mathbf{U}]^{-1}_{ij}\right)^2\right\}^{1/2} \Big/ 6, \qquad (11.4.3.8)$$

where i and j are the indices of the 3 × 3 matrix $[\mathbf{U}]$.

---

[2] This is when the crystal oscillation angle during the measurement of a single diffraction pattern is larger than the angular reflection width.

The value of this index increases as additional symmetry restraints are imposed, starting from zero for a triclinic cell. Autoindexing in *DENZO* always finishes with a table of distortion indices for the 14 possible Bravais lattices, but does not automatically make a choice of lattice.

11.4.3.2. Lattice pseudosymmetry

The cell-reduction procedure cannot determine lattice symmetry, since it cannot distinguish true lattice symmetry from a lattice accidentally having higher symmetry within experimental error (*e.g.* a monoclinic lattice with $\beta \simeq 90^\circ$ is approximately orthorhombic). If one is not certain about the lattice symmetry, the safe choice is to assume space group $P1$, with a primitive triclinic lattice for the crystal, and to check the table again after refinement of the diffraction-geometry parameters. A reliable symmetry analysis can be done only by comparing intensities of symmetry-related reflections, which is done later in *SCALEPACK* or another scaling/merging program.

11.4.3.3. Data-collection requirements

The total oscillation range needed for autoindexing has to cover a sufficient number of spots to establish the periodicity of the diffraction pattern in three dimensions. It is also important that the oscillation range of each image is small enough so that the lunes (*i.e.* rings of spots from one reciprocal plane) are resolved. One should note that the requirement for lune separation is distinct from the requirement for spot separation. If lunes overlap, spots may have more than one index consistent with a particular position on the detector.

The autoindexing procedure described above is not dependent on prior knowledge of the crystal unit cell; however, for efficiency reasons, the search is restricted to a reasonable range of unit-cell dimensions, obtained, for example, from the requirement of spot separation. In *DENZO*, this default can be overridden by the keyword *longest vector*, but the need to use this keyword is a sign of a problem that should be fixed. Either the defined spot size should be decreased or data should be recollected with the detector further away from the crystal.

11.4.3.4. Misindexing

Autoindexing is sensitive to inaccuracy in the description of the detector geometry. The specified position of the beam on the detector should correspond to the origin of the Bragg-peaks lattice (*i.e.* the peak with Miller index 000). Autoindexing will shift the origin of the lattice to the Bragg lattice point closest to the specified beam position. An incorrect specification of the beam position will result in the incorrect placement of the Miller index 000. In such a situation, all reflections will have incorrectly determined indices. Such misindexing can be totally self-consistent until the intensities of symmetry-related reflections are compared. This dependence of the indexing correctness on the assumed beam position is one of the main sources of difficulties in indexing (Otwinowski & Minor, 1997; Gewirth, 2003). The beam position has to be precise, as the largest acceptable error is one half of the shortest distance between spots.

The process of determining $\{h, k, l\}$ triplets is not very sensitive to other detector parameters. Errors of a degree or two in rotation or by 10% in distance are unlikely to produce wrong values of h, k and l. Sometimes data with even very large errors, *e.g.* if the detector distance is too large by a factor of 5, will still produce the correct $\{h, k, l\}$ triplets. The detector position error is compensated by an error in the lattice determined by autoindexing. For this reason, the accuracy of the lattice is not a function of the autoindexing procedure, but depends mainly on the accuracy of the detector description. By the same token, the distortion of the lattice also depends on the accuracy of the detector parameters.

11.4.3.5. Twins

Special care has to be taken if more than one crystal contributes to the diffraction image. When there is a large disproportion between the volumes of the crystals (*e.g.* the presence of a satellite crystal), autoindexing may work without any modifications. In the case of crystals of similar volumes, manual removal of weaker reflections from the peak-search list and resolution cuts can make the proportion of reflections from one crystal in the peak list large enough for the autoindexing method to succeed. If the multiple crystals have similar orientations, using only very low resolution data may be the right method. In the case of twinned crystals, the autoindexing procedure sometimes finds a superlattice that simultaneously assigns integer indices to reflections from both crystals. In such a case, *DENZO* solves the problem of finding the best three-dimensional lattice that incorporates all of the observed peaks. Unfortunately, for a twinned crystal, the periodicity of the overall diffraction pattern may not in all cases be the proper criterion to index two different lattices. Alternative approaches have been described in the literature for indexing such cases.

11.4.4. Coordinate systems

There are four natural coordinate systems used to describe a diffraction experiment, defined by the order in which the diffraction-image pixels are stored in the detector, the beam and gravity, or the beam and the goniostat axis (spindle or 2θ). These coordinate systems will be called *data*, *beam–gravity*, *beam–spindle* and *beam–*2θ, respectively.

11.4.4.1. Beam–gravity

To visualize a diffraction pattern, beam–gravity is the coordinate system clearly preferred by human physiology. The universal preference to relate to the gravity direction is revealed by the observation that people generally perceive an image in a mirror as inverted left–right rather than top–down. Hence *XdisplayF* uses the beam–gravity coordinate system.[3]

11.4.4.2. Data

The first (1983) *DENZO* implementation used the data coordinate system to describe the beam position on the detector and to define the integration box. This is still the case in order to keep backward compatibility. There are eight ways of relating detector-data order to the beam–gravity coordinate system. All of them have been encountered in various detector formats and are the part of the detector-format description in *DENZO*.

11.4.4.3. Beam–spindle

Initially, *DENZO* supported only a single-axis goniostat and used a beam–spindle coordinate system to define the crystal and detector orientation, as well as polarization. The goniostat spindle axis was assumed to be horizontal, so the direction perpendicular to the beam and spindle was described by the keyword *vertical*, which in reality may not relate to the gravity direction for some goniostats. The keyword *rotx* relates to rotation around the *spindle axis*, *roty* around the *vertical axis* and *rotz* around the *beam axis*. The definition of the orientation matrix in the file used for communication between *DENZO* and *SCALEPACK* uses an unintuitive convention: the letter *y* in *roty* relates to the first element of the vector, *x* in *rotx* to the second and *z* in *rotz* to the third. However, the matrix always has a positive determinant, so this convention has no impact on the handedness of the coordinate system. This unfortunate choice of convention, preserved for backward compatibility reasons, appears only in the communication file and has no significance for anybody who does not inspect the matrix.

---

[3] There may occasionally be an exception to this when the experimental system is not known.

11.4.4.4. Beam–2θ

The description of generalized multiple-axis goniostats introduced a conceptual change in the *DENZO* coordinate system. The data-collection axis can be oriented in any direction, so in principle *rotx*, *roty* and *rotz* no longer need to be defined relative to the data-collection axis. However, to keep the useful correlations between refinable parameters (for example, *crystal rotz* and *detector rotz* are typically close to 100% correlated), one real and two virtual goniostats are used simultaneously in *DENZO*. Refinable crystal parameters (*crystal rotx*, *roty*, *rotz*) are still defined, as in the past, by the data-collection axis and the beam. This means that the directions of rotations defined by *fit crystal rotx*, *roty* and *rotz* do not rotate around the data-collection axis as the program advances from one image to another. This coordinate system changes with the change in direction of the data-collection axis. The crystal orientation is defined by three constant, perpendicular axes, which do not have to be aligned with the physical crystal goniostat. However, the so-called *2 theta* rotation has a fixed axis, and, if it exists, it defines the *DENZO* coordinate system together with the beam axis. Thus, the current coordinate system in *DENZO* should be called beam–2θ. Fortunately for the user, the conversions between different coordinate systems are handled transparently. For example, the refined change in the crystal orientation is converted from the refined axes goniostat to the crystal-orientation goniostat. The movements of the physical goniostat are converted into appropriate changes in the diffraction pattern. The physical goniostat appears only to describe the data collection and, optionally, to calculate the physical goniostat angles needed to produce particular crystal alignments.

The *DENZO* coordinate system (Gewirth, 2003) is used in the definition of crystal goniostats, 2θ goniostat and polarization. The curvature of cylindrical detectors and Weissenberg coupling are described in this coordinate system as well.

This discussion of the coordinate systems shows that the conceptual complexity of the program description can be compatible with – or even lead to – simplicity in the use of the program. The success of data analysis does not require a full understanding of the relations between internal *DENZO* goniostats and the coordinate systems when using the programs. The reason for this complexity was to create a simple pattern of correlations between crystal and detector parameters in *DENZO* refinement. This in turn allows for simple and easy-to-understand control of the refinement process and simplifies problem diagnostics. For example: the definition of refined *crystal rotx* as rotation around the data-collection axis makes the spindle and shutter problems manifest only as fluctuations of *crystal rotx*. Constant nonzero values of refined shifts between frames of *crystal roty* and *rotz* are a sign of misalignment of the data-collection axis. Although the program compensates for this misalignment with changes in crystal orientation, this introduces a small error in the Lorentz factor (which will be, to a good approximation, compensated for in scaling). The nature of these problems is such that they do not result in a complete failure of the experiment, but they do have an impact on the quality of the result. It is up to the experimenter and the instrument manager to assess the significance of these indications.

11.4.5. Experimental assumptions

To achieve the main goal of a diffraction experiment – the estimation of structure-factor amplitudes – three components need to be determined, with highest possible precision:

(1) the crystal response function, namely the relationship between the crystal structure factor and the number of diffracted X-ray photons;
(2) the detector response function; and
(3) the geometrical description of the detector relative to the directions of the X-ray beam and crystal goniostat axes.

The main difficulty of data analysis in protein crystallography is the complexity of the process that determines these components. *HKL* can determine all three directly from the data produced by the analogue-to-digital converter (ADC). The only extra program needed is one that sends the raw ADC signal to the computer disk. For charge-coupled-device (CCD) detectors, spatial detector distortion and sensitivity per pixel functions need to be established in a separate experiment. An important part of such calibration is the flood-field exposure, which can provide a pixel-by-pixel response correction for calculating the background in the X-ray exposure. There was hope that it would also correct for localized variations in response to Bragg-peak intensities, but it has been shown that light diffusion in reducing fibre optic taper depends on the stretching process during individual manufacture of such tapers. Consequently, there are substantial effects, in addition to the flood-field exposure, on Bragg-peak intensities towards the edges of such tapers. For single taper (single CCD) detectors, with the beam in the centre, the Bragg peaks are underestimated towards the edge in a resolution-dependent way, but without noticeably affecting merging statistics. The resolution-dependent effect is mostly compensated for by a slight adjustment in temperature factor of the model, so it has little practical impact. The issue is much more severe for larger, multi-taper (multi-CCD) detectors, where the localized effects may greatly affect the phasing signal. This effect can be corrected to a degree during scaling, in which geometric details of the detector construction are taken into account.

Additionally, determination of detector-geometry distortion can be accomplished with diffraction experiments on a well diffracting, high-symmetry, non-slipping crystal and a special data-collection procedure. Such a procedure can be used for experimental detectors or to correct for manufacturers' inaccuracies, which are occasionally noticeable.

In practice, many manufacturers of CCD detectors provide post-processed data rather than ADC output, to hide features and flaws of their instruments. This practice complicates data analysis, as the post-processing does not correct well for variations in CCD detector sensitivity, changes the shapes of diffraction-peak profiles and creates incorrect estimates of diffraction-peak intensities in places where flaws were hidden.

Better results would be obtained by allowing output of data directly from the ADC, letting CCD detectors approach the lower level of systematic errors observed for image-plate detectors.

11.4.5.1. Crystal diffraction

The crystal response function consists of two types of factors, which are included in the analysis: additive factors represented by the background, and a number of multiplicative factors, such as exposed crystal volume, resolution-dependent decay, Lorentz factor, flux variation, absorption, polarization *etc.* Other factors, like extinction, are currently ignored by *HKL*, except for their contribution to error estimates. Specific radiation-induced damage (Borek *et al.*, 2007, 2010) is approached differently, as it is a merging, rather than scaling-factor, correction.

11.4.5.2. Data model

The detector response function is one of the main components of the data model. The detector formats define the coordinate system, the pixel size, the detector size, the active area and the fundamental shape (cylindrical, spherical, flat rectangular or circular, single or multi-module) of the detector.

HKL supports most of the data formats representing particular combinations of the features listed below:

1. data stored in 8 or 16 bit fields;
2. overflow table;
3. linear, bilinear, polynomial and exponential response, with the error model represented by an arbitrary scale and offset;
4. saturation limit;
5. value representing lack of data;
6. constant offsets per read-out channel;
7. pattern noise;
8. lossless compression;
9. flood-field response; and
10. sensitivity response.

The main complexity of the data-analysis program and the difficulties in using it are not in the application of the data model but rather in the determination of the unknown data-model parameters. The refinement of the data-model parameters is an order of magnitude more complex in terms of the computer code than the integration of Bragg peaks when the parameters are known.

The data model is a compromise between an attempt to describe the measurement process precisely and the ability to find parameters describing this process. For example, the overlap between Bragg peaks is typically ignored due to the complexity of spot-shape determination when reflections overlap. The issue is not only to implement the parameterization, but also to do it with acceptable speed and stability of the numerical algorithms. A more complex data model can be more precise under specific circumstances, but can result in a less stable refinement and produce less precise final results in most cases. An apparently more realistic but also more complex data model may end up being inferior to a simpler and more robust approach. The complexity of model-quality analysis is due to the fact that some types of errors may be much less significant than others. In particular, an error that changes the intensities of all reflections by the same factor only changes the overall scale factor between the data and the atomic model. Truncation of the integration area results in a systematic reduction of calculated reflection intensities. A variable integration area may result in a different fraction of a reflection being omitted for different reflections. The goal of an integration method is to minimize the variation in the omitted fraction, rather than its magnitude. Similarly, if there is an error in predicting reflection-profile shape, a constant error has a smaller impact than a variable error of the same magnitude.

The magnitudes and types of errors are very different in different experiments. The compensation of errors also differs between experiments, making it hard to generalize about an optimal approach to data analysis when the data do not fully satisfy the assumptions of the data model. For intense reflections, when counting statistics are not a limiting factor, none of the current data models accounts for all reproducible errors in experiments. This issue is critical in measuring small differences originating from dispersive effects, *i.e.* anomalous scattering.

11.4.5.3. Data-model refinement

The parameters of the data model can be classified into four groups:

1. Those determined and refined from self-consistency of the data by a (nonlinear) least-squares method.
2. Parameters that can be determined from internal self-consistency of the data, but for which least squares are not implemented.
3. Parameters that have to be established in a separate experiment, *e.g.* pixel sensitivity from flood-field exposure.
4. Parameters that are obtained from a description of the hardware.

The least-squares method is based on minimization of a function that is a sum of contributors of the following type:

$$(\mathrm{pred} - \mathrm{obs})^2/\sigma^2 = \chi^2, \qquad (11.4.5.1)$$

where pred is a prediction based on some parameterized model, obs is the value of this prediction's measurement and σ^2 is a combined estimate of the measurement and the prediction uncertainties (variance). *DENZO* refines the following parameters by least-squares methods:

1. unit-cell vectors in autoindexing;
2. reflection background and background slope; and
3. crystal orientation, unit cell, mosaicity, beam focus and position, detector orientation and position, and geometrical distortions that are parameterized differently for different types of detectors.

SCALEPACK can refine the following parameters by least-squares methods:

1. unit cell, crystal orientation and mosaicity, including changes of these parameters during an experiment;
2. goniostat internal alignment angles;
3. crystal absorption, using a spherical-harmonic expansion (Katayama, 1986; Blessing, 1995) of the absorption surface;
4. uniformity of exposure, including shutter-timing error;
5. correction to the Lorentz factor, resulting from misalignment of the spindle axis;
6. reproducible wobble of the rotation axis, resulting from misalignment of gears in a spindle assembly;
7. non-uniform smooth detector response, resulting (for example) from light diffusion in the fibre optic taper not correctable by flood-field exposure; and
8. other factors contributing to scaling, such as slow fluctuations of beam intensity, changes in exposed volume of the crystal, overall crystal decay and resolution-dependent crystal decay.

SCALEPACK can also refine, by Tikhonov-stabilized least-squares methods (Borek *et al.*, 2010), multiple structure-factor components contributing to a group of reflections with the same unique index:

(1) the zero-dose extrapolated intensity of a unique reflection;
(2) perturbations resulting from Friedel-symmetry violations (the so-called anomalous signal);
(3) perturbations resulting from wavelength-dependent changes (the so-called dispersive f' differences);
(4) perturbations resulting from radiation-induced specific changes of structure factors;
(5) other changes in structure, resulting from differences between crystals or pseudosymmetry (imperfect rotational symmetry);
(6) higher-order terms of all the above-mentioned effects.

11.4.5.4. Correlation between parameters

Occasionally, the refinement can be unstable due to high correlation between some parameters. Refinement of highly correlated parameters, for example positional refinement of unit-cell parameters and the distance between the crystal and the detector, results in the errors in one parameter compensating for errors in other parameters. However, such error compensation may still lead to mostly correct predictions; for example, incorrect distance and unit-cell parameters can result in correct predictions of the diffraction pattern. In an extreme case, where compensation is 100%, the parameters involved would be undefined, but the error compensation by other parameters involved would be undefined even when the diffraction-pattern prediction is successful.

To prevent instability in calculations, eigenvalue filtering (Reeke, 1984) is employed to remove the most correlated components from the refinement. Eigenvalue filtering works reliably when the starting parameters are close to the correct values. It may fail, however, to correct large errors in the input parameters if the correlation between parameters is close to 100%. All refinements have also built-in Tikhonov or Tikhonov-like stabilizers (Tikhonov & Arsenin, 1977) to prevent unreasonable values of parameters or differences between them.

Once the whole data set is integrated, post refinement (Rossmann, 1979; Winkler *et al.*, 1979; Evans, 1987; Greenhough, 1987; Kabsch, 1993) can refine unit-cell parameters and crystal-orientation parameters more precisely and without correlation with detector parameters. For this reason, the unit-cell parameters used in structure-determination calculations should come from the post refinement performed in *SCALEPACK* and not from initial refinement in *DENZO*.

11.4.5.5. Single- and multiframe refinement

The crystal and detector orientation parameters can be refined either for a group of images or for each processed image separately. Refinement performed separately for each image allows for robust data processing, even when the crystal slips considerably during data collection.

11.4.5.6. Active area

Not every pixel represents a valid measurement. Specification of the active detector area in *DENZO* is derived from the format and the definition of the detector size. Detector calibration with flood-field exposure will calculate the sensitivity for each pixel and will also determine which pixels should be ignored. Unfortunately, rather than ignoring bad areas, some manufacturers enter values interpolated from nearby pixels, in particular to present a fake contiguous image for multi-module detectors. This practice results in observations in these areas not scaling well with their symmetry equivalents.

The input command in *DENZO* can additionally label some areas of the detector to be ignored, most frequently the shadow caused by the beam stop and its support. There are also commands to ignore triangular shapes, margins of the detector and a particular line or pixel.

11.4.5.7. Flood field

The basic method for calibration of the spatial dependence of detector sensitivity is to measure the response of the detector to a flood-field exposure. The amount of relative exposure per pixel needs to be known. *DENZO* allows for either a uniform or an isotropic source. If the source is at the crystal position, *DENZO* refinement (with a separate crystal exposure) can be used to define the geometry of the source relative to the detector. To calculate the flood-field response, a prior determination of the detector distortion is required. The flood-field response is converted to a sensitivity function. Large deviations from the local average are used to define inactive pixels. The edge of the active area needs special treatment, dependent on the method of phosphorus deposition.

11.4.5.8. Absolute configuration

Absolute configuration is defined relative to the data-coordinate system and is only affected by the sign of the parameter *y scale*. A combined mirror transformation of the image data and of the goniostat description does not affect the self-consistency of the data. Thus, the correctness of the absolute configuration cannot be verified by data-reduction programs. Owing to inherent ambiguities in indexing when merging data from different crystals, re-indexing is sometimes necessary, which may cause unjustified concern that the re-indexing may somehow affect the sign of Bijvoet differences. *SCALEPACK* checks for the positive sign of the re-indexing matrix, and thus the procedure cannot affect the absolute configuration or the Bijvoet differences.

11.4.5.9. Detector goniostat

The detector goniostat in *DENZO* can have only one rotation axis – 2θ. In the complex transformations described in equation (11.4.2.8), the geometrical scale is affected by pixel-to-millimetre conversion and distortion. For different instruments, the scale of the rotation is defined differently. For detectors without distortion, the scale is defined by the value of the pixel size in the 'slow' direction. For detectors with distortion characterized by polynomials (*e.g.* CCD detectors), the scale is defined in the way the distortion was determined. In such a case, the scale is derived either from the separation between holes in the reference grid mask or the detector translation. As the distance of the detector active surface from the crystal cannot be measured precisely, in the latter approach the difference between the two distances is the ultimate source of the scale reference. The angle between the detector distance translation and the X-ray beam completes the definition of the detector goniostat in *HKL*.

11.4.5.10. Crystal goniostat

Each type of physical goniostat is defined by six angles. Two angles define the direction of the main axis (ω) in the *DENZO* coordinate system. The third angle defines the zero position of the ω axis. The fourth is the angle between ω and the second axis (κ or χ). The fifth defines the zero position of the second axis. The sixth is the angle between the second and the third axes. This type of goniostat definition allows for the specification of any three-axis goniostat (EEC Cooperative Workshop on Position-Sensitive Detector Software, 1986). Misalignment of the goniostat is represented as an adjustment to these angles, which can be refined by the *HKL* system.

11.4.5.11. Crystal orthogonalization convention

Crystal orientation specified by the three angles needs a definition of a zero point. Any crystal axis, or the equivalent reciprocal-space zone perpendicular to it, can be used as a reference. The definition of zero point aligns the crystal axis with the beam direction and one of the reciprocal axes with the x direction in the *DENZO* coordinate system. The user can specify which crystal axes will be aligned at zero crystal orientation angles by specifying the so-called reference zone.

11.4.5.12. Refinement and calibration

Both the refinement and calibration procedures determine the properties of the instrument. The principal difference between refinement and calibration is that calibration is performed with data not obtained during the current diffraction experiment, and refinement uses data obtained during the current diffraction experiment. *DENZO* performs both refinement and calibration. In some cases, the difference between calibration and refinement is a question of semantics, as the refined data from one experiment can be used as a reference for another experiment, or even as a reference for a subsequent refinement cycle or for another part of the same experiment.

11.4.6. Prediction of the diffraction pattern

The autoindexing procedure assigns Miller indices only to strong spots, ones that can be found through a peak-search procedure. The target of the experiment is to estimate structure-factor amplitudes for all reflections captured by the detector. Therefore, positions of all spots need to be predicted by applying the following equations to all possible triplets $\mathbf{h}$. Using

$$\mathbf{S} = [\mathbf{A}]^{-1}\mathbf{h}, \qquad (11.4.6.1)$$

a matrix $[\mathbf{A}]$ must be found that generates the vector $\mathbf{S}$ that satisfies the diffraction condition [equation (11.4.2.1)], knowing that the matrix $[\mathbf{A}]$ is a function of the crystal orientation at the goniostat angles that generated the reflection with indices $\mathbf{h}$ [equation (11.4.2.6)]. The rotation of the crystal during the experiment creates a straightforward algebraic problem that results in a complex equation defining the angle at which the reflection occurs. This angle consequently defines the image or images on which the reflection appears. Knowing this angle, the vector $\mathbf{S}$ can be calculated, and, from equation (11.4.2.5), the direction of the vector $\mathbf{X}$ can be found:

$$\mathbf{X}/|\mathbf{X}| = \lambda(\mathbf{S} - \mathbf{S}_0). \qquad (11.4.6.2)$$

Calculation of the length of vector $\mathbf{X}$ requires a knowledge of detector orientation, which, for flat detectors, is described here by vector $\mathbf{G}$, perpendicular to the detector and with length equal to the crystal-to-detector distance:

$$\mathbf{Y} = [\mathbf{R}_{2\theta}]^{-1}\lambda(\mathbf{S} - \mathbf{S}_0), \qquad (11.4.6.3)$$

$$\mathbf{H} = [\mathbf{R}_z]^{-1}[\mathbf{R}_y]^{-1}[\mathbf{R}_x]^{-1}\mathbf{G}, \qquad (11.4.6.4)$$

$$\mathbf{X} = [\mathbf{R}_{2\theta}]\frac{\mathbf{H}\cdot\mathbf{H}}{\mathbf{Y}\cdot\mathbf{H}}\mathbf{Y}. \qquad (11.4.6.5)$$

Then, by inverting equation (11.4.2.8), the position in pixels $\{p, q\}$ of the reflection can be calculated:

$$\begin{aligned}\{p, q\} = {} & [\mathbf{D}]^{-1}[\mathbf{K}]^{-1}([\mathbf{L}]^{-1}([\mathbf{R}_{2\theta}]^{-1}([\mathbf{R}_z]^{-1}[\mathbf{R}_y]^{-1}[\mathbf{R}_x]^{-1}(\mathbf{X} - \mathbf{T}_D) \\ & + \mathbf{T}_D) - \mathbf{T}_D) + \mathbf{B}).\end{aligned} \qquad (11.4.6.6)$$

11.4.6.1. Refinement of crystal and detector parameters

The precision of the integration step depends on precise knowledge of the peak positions. The autoindexing step provides only an approximate orientation of the crystal, and the result of that step is imprecise if the initial values of the detector parameters are poorly known. A nonlinear least-squares refinement process is used to improve the prediction (EEC Cooperative Workshop on Position-Sensitive Detector Software, 1986). Depending on the particulars of the experiment, the same parameters (*e.g.* crystal-to-detector distance) may either be known more precisely *a priori*, or are better estimated from the diffraction data. *DENZO* allows either fixing or refining of each of the parameters separately. This flexibility is important when characterizing a detector, but when detector parameters are already known, the *fit all* option and detector-specific default values are quite reliable.

DENZO can refine the six parameters describing the position and orientation of the detector in space. It can also refine internal parameters of the detector including:

(1) curvature radius and rotation of the active surface – *film rotation*, for cylindrical detectors;
(2) x to y scale and x to y skew;
(3) radial/angular distortion for spiral scanners; and
(4) polynomial distortion, separate for each CCD module of multi-CCD detectors (Naday *et al.*, 1998).

Detector- and crystal-parameter refinement in *DENZO* is achieved by minimizing the sum of three functions [equations (11.4.6.7), (11.4.6.8) and (11.4.6.11)] of the type in equation (11.4.5.1). The contribution resulting from the measurement of position $\{p, q\}$ of the reflection is

$$\chi_p^2 = \sum_{hkl}(p_{\rm pred} - p_{\rm cent})^2/\sigma_p^2, \qquad (11.4.6.7)$$

$$\chi_q^2 = \sum_{hkl}(q_{\rm pred} - q_{\rm cent})^2/\sigma_q^2. \qquad (11.4.6.8)$$

where $p_{\rm pred}, q_{\rm pred}$ is the predicted position of the reflection in pixel coordinates, $p_{\rm cent}, q_{\rm cent}$ is the centroid position of the observed reflection in pixel coordinates and σ_p, σ_q are combined estimates of uncertainties of the observed and predicted positions.

11.4.6.2. Bragg's law for non-ideal conditions: mosaicity

The Bragg condition [equation (11.4.2.1)] assumes diffraction from ideal crystals and a parallel X-ray beam. In reality, crystals are mosaic and the beam has some angular spread. The value of the *mosaicity* keyword describes the range of orientations of the crystal lattice within a sample. As the impacts of mosaicity and the beam's angular spread on the angular width of reflections are equivalent, the keyword *mosaicity* describes the sum of both effects.

DENZO assumes the following model of angular shape of diffraction peaks for the oscillation angle φ:

$$M = \frac{\text{mos}}{\pi}\left[1 + \cos\frac{\pi(\varphi - \varphi_c)}{\text{mos}}\right], \qquad (11.4.6.9)$$

where mos is the observed angular width of a diffraction peak yet to be increased by the Lorentz factor, φ_c is the predicted oscillation angle at which the diffraction condition is fulfilled, φ is in the range($\varphi_c - \text{mos}/2$; $\varphi_c + \text{mos}/2$), otherwise $M = 0$. The actual width of the reflection is different due to Lorentz-factor variability over the image, so equation (11.4.6.9) describes only the common component of the angular width.

Using equation (11.4.6.9) we can calculate

$$P_{\varphi_1\varphi_2} = \int\limits_{\varphi_1}^{\varphi_2} M(\varphi)\,\mathrm{d}\varphi. \qquad (11.4.6.10)$$

P is the predicted partiality of data collected by oscillating from φ_1 to φ_2. It is a number that represents what fraction of the reflection intensity is present in one image. If the partiality is 1, such reflections are called fully recorded; otherwise, they are called partials. For partials, predictions of partiality can be compared with the observed fraction P_0 of the reflection intensity present in one image. The partiality model contributes the following term to the refinement:

$$\chi^2_P = (P_{\varphi_1\varphi_2} - P_0)^2/\sigma^2_{P_0}. \qquad (11.4.6.11)$$

The combined positional [described by equations (11.4.6.7) and (11.4.6.8)] and partiality refinement [equation (11.4.6.11)] used in *DENZO* is both stable and very accurate. The power of this method lies in proper weighting by estimated errors of two very different terms – one describing positional differences and the other describing intensity differences. Both detector and crystal variables are uniformly treated in the refinement process.

11.4.6.3. Detector distortions

The design of detectors results in pixels not being positioned on an exact square or rectangular grid. A correct understanding of the detector distortions is essential to accurate positional refinement. The types of distortions are detector-specific. The primary sources of error include misalignment of the detector position sensors and optical or magnetic distortion in CCD-based detectors. If the detector distortion can be parameterized, then these parameters should be added to the refinement. For example, in the case of spiral scanners, there are two parameters describing the end position of the scanning head. In a perfectly adjusted scanner, these parameters would be zero. In practice, however, they may deviate from zero by as much as 1 mm. Such misalignment parameters can correlate very strongly with other detector and crystal parameters, particularly for low-symmetry lattices or for low-resolution data. If the distortions are stable, it is better to determine them in a separate experiment optimized for that task.

Fibre-optic tapers used in many CCD detectors have distortions that have to be individually determined for each instrument. The distortion is stable over time and its spatial characteristics are dominated by a smooth component and a small local shear. In high-quality tapers used in X-ray instruments, the small local shear can be ignored. The smooth component can be parameterized in a number of ways, for example by splines or polynomials (Messerschmidt & Pflugrath, 1987). *DENZO* uses two-dimensional Chebyschev polynomials (Press *et al.*, 1989) in $\{x, y\}$ or $\{p, q\}$ coordinates, normalized to the range $\{[-1, +1], [-1, +1]\}$. Typically, fifth- or seventh-order polynomials result in a positional error (r.m.s.) lower than 7 µm, which is about one tenth of a typical detector pixel. *DENZO* can use either a grid mask pattern or the X-ray diffraction pattern to refine the coefficients of the Chebyschev polynomials. If a grid mask is used, it has to be precisely made and positioned. The use of crystallographic data requires precise knowledge of detector and crystal parameters that are not known *a priori* with the required precision. The crystal and detector parameters can be determined in the same experiment as detector distortion. However, this experiment needs to be designed to minimize the impact of correlations between the parameters involved. The data analysis requires the description of the distortion function and its inverse. In *DENZO*, both are approximated in terms of Chebyschev polynomials. The magnitude of the approximation error is the same for the distortion function and its inverse.

11.4.7. Integration of diffraction maxima by profile fitting

To accurately integrate diffraction peaks, the spot position has to be predicted accurately. Each integration program has its own procedure for predicting and fitting diffraction profiles. In *DENZO*, profile-shape prediction is defined by a weighted average of other reflection profiles present within some radius from the spot of interest. Each spot has its own prediction, which is continuously adjusted to variations of spot shapes over the detector. The profiles are added by shifting them to the same position, generating a normalized profile P_i, where $\sum_i P_i = 1$. In the second step, the measured pixels' values are fitted to a function

$$B_i + IP_i, \qquad (11.4.7.1)$$

where B_i is the predicted value for the background in pixel i and I is the diffraction intensity of the spot. The profile-fitting procedure minimizes the function

$$\sum_i \frac{[M_i - (B_i + IP_i)]^2}{V_i}, \qquad (11.4.7.2)$$

where M_i are the measured pixel values and V_i are the variances of these measurements. The minimum of this function defines the value of the profile-fitted intensity (Otwinowski & Minor, 1997):

$$I = \frac{\sum_i P_i(M_i - B_i)/V_i}{\sum_i P_i^2/V_i}. \qquad (11.4.7.3)$$

11.4.8. Scaling – multiplicative corrections

The relation between the measured intensity $I(hkl)$ of reflection hkl and its squared structure-factor amplitude $|\mathbf{F}(hkl)|^2$ is described by

$$I(hkl) = I_b r_e^2 \frac{\lambda^2}{|\mathbf{S} \times \boldsymbol{\omega}'|} PT \frac{V}{v_u^2} |\mathbf{F}(hkl)|^2 D_A D_S, \qquad (11.4.8.1)$$

where I_b is the flux density of the primary beam; $r_e = e^2/m_e c^2$ is the classical electron radius (2.818×10^{-12} mm); λ is the wavelength of the beam; $\mathbf{S} \times \boldsymbol{\omega}'$ is a cross product between the diffraction vector $\mathbf{S}$ and $\boldsymbol{\omega}'$ (the projection of the crystal-rotation-speed vector $\boldsymbol{\omega}$ on the plane perpendicular to the primary beam;

thus $|\boldsymbol{\omega}|/|\mathbf{S}\times\boldsymbol{\omega}'|$ is the Lorentz factor); P is a polarization factor (Azaroff, 1955); T is the transmission of the beam [related to the absorbance defined as $A = -\ln T$, $T = \exp(-A)$]; v_u is the volume of a primitive crystal unit cell; V is the volume of the crystal exposed to the beam; $|\mathbf{F}(hkl)|^2$ is the square of the structure-factor amplitude for the given reflection hkl; D_A is the absorption of X-rays by the detector's active material; and D_C is the detector's response to a single absorbed X-ray photon.

SCALEPACK determines the components of the total scale factor, *i.e.* the product of all factors that multiply the structure-factor amplitudes squared in equation (11.4.8.1):

$$I(hkl) = K|\mathbf{F}(hkl)|^2, \qquad (11.4.8.2)$$

where I is the intensity and K is the total scale factor.

All components of the total scale factor K can be calculated from non-diffraction measurements and calibration of the data-collection system. However, the absolute calibration of the whole system is rarely available and part of the scale factor (the overall scale factor k_o) is determined by comparing the scaled data to the squared structure-factor amplitudes predicted from an atomic model obtained after the structure is solved:

$$K = k_o k_r = k_o(k_{\rm beam}\, k_{\rm polarization}\, k_{\rm detector}\ldots). \qquad (11.4.8.3)$$

The relative scale factor k_r is calculated, but in practice we assume that some of its components are known from detector calibration, beam monitoring and the diffraction geometry. The scaling procedure determines the remaining parts of the scale factor (often iteratively), based on knowledge from subsequent stages of crystallographic analysis. When iterative procedures are applied, the scaling model has to include information about experimental uncertainties (discussed in Section 11.4.10).

SCALEPACK uses an exponential modelling approach (Otwinowski *et al.*, 2003), which is flexible with regard to correlations among parameters and provides a uniform description of the parameter optimization process for various scaling models. In *SCALEPACK* scale factors s_i for each observation are calculated using a set of *a priori* unknown parameters p_i:

$$s_i = \exp\left[\sum_i p_i\, f_i(hkl)\right], \qquad (11.4.8.4)$$

where f_i are pre-defined modelling functions of experimental conditions and i is a hierarchical index referring both to the type of correction and to the indices of the functional parameters describing this correction.

The simplest scaling model has a separate scale factor for each group (or batch) of data, *e.g.* one scale factor per image. In such a case,

$$f_{i_s} = \delta_{ij}, \qquad (11.4.8.5)$$

where j is the batch index for a particular reflection. From equation (11.4.8.4) we obtain P_{i_s} as the logarithm of the scale factor of batch i.

To correct for a radiation-damage component represented as resolution-dependent decay, one temperature factor is used per batch of data,

$$f_{i_B} = [(\mathbf{S}\cdot\mathbf{S})/2]\delta_{ij}, \qquad (11.4.8.6)$$

where $\mathbf{S}$ is the scattering vector for each reflection. Using the same approach as for equation (11.4.8.5), p_{i_B} is obtained as the relative temperature factor.

Another, more complex, multiplicative correction addresses unknown crystal absorption using an average of absorbances in the incoming and diffracted beam directions (Kopfmann & Huber, 1968). The correction is parameterized by real spherical harmonics as a function of the direction of the incoming beam and the diffraction vector $\mathbf{S}$, expressed in the polar coordinate system of the rotating crystal (Katayama, 1986; Blessing, 1995):

$$\begin{aligned} f_{as,lm} &= \frac{1}{2}\left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right]^{1/2}[P_{lm}(\cos\theta_i)\sin(2\pi m\Phi_i) \\ &\quad + P_{lm}(\cos\theta_o)\sin(2\pi m\Phi_o)], \\ f_{ac,lm} &= \frac{1}{2}\left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!}\right]^{1/2}[P_{lm}(\cos\theta_i)\cos(2\pi m\Phi_i) \\ &\quad + P_{lm}(\cos\theta_o)\cos(2\pi m\Phi_o)], \end{aligned} \qquad (11.4.8.7)$$

where as, lm and ac, lm are parts of the hierarchical index i from equation (11.4.8.4); l, m are indices of the spherical harmonics; P_{lm} is a Legendre polynomial; and θ, Φ are the polar coordinates of the incoming (index i) and the outgoing (index o) directions in the crystal coordinate system. The odd-order spherical harmonics should have zero coefficients when describing pure absorption of X-rays. However, due to correlation with other effects, the scaling may benefit from the inclusion of low-order odd-order harmonics.

It is beneficial to correct even for a small discrepancy between the actual and assumed directions of the crystal rotation axis. The inaccuracy results in an error in the calculated value of the Lorentz factor [equation (11.4.8.1)]. The scale factor to correct for this error can be described using the parameter p_l, the value of which represents a small angular error. The corresponding function is

$$f_l = \frac{\mathbf{S}\cdot\boldsymbol{\omega}'}{|\mathbf{S}\times\boldsymbol{\omega}'|}. \qquad (11.4.8.8)$$

There are additional effects that are parameterized using exponential modelling [equation (11.4.8.4)], for example, a correction for uneven crystal rotation and/or exposure (Otwinowski *et al.*, 2003), or a correction for uneven detector response, which is very important for multi-CCD detectors. This approach can be extended to other experimental factors if there is a need to correct them during scaling without changing the overall logic of scale-factor determination (Otwinowski *et al.*, 2003).

11.4.8.1. Global and local scaling

In principle, global scaling can be followed by local scaling (Matthews & Czerwinski, 1975). Local scaling is mostly applied to calculate differences of phasing signal, where it is assumed that a group of measurements, *e.g.* those close together in reciprocal space or detector space, should be on a similar scale. A flexible parameterization by the exponential modelling allows for a good description of all kinds of smooth corrections. Local scaling is much more limited in terms of what type of smooth variation is being corrected for, so it is unlikely to provide additional benefit to the general scaling method described here. In practice, if there is an improvement from such procedures at the stage of heavy-atom search, it implies that the scaling parameters for global scaling were not properly chosen.

11.4.8.2. Stabilization of scaling parameters based on prior knowledge

The unknown parameters in equation (11.4.8.4) are estimated with various level of uncertainty depending on the multiplicity of observations and how symmetry-equivalent reflections are related to each other. Potentially, this may result in unreasonable values of scaling parameters due to insufficient information to determine the values of parameters. In *SCALEPACK*, the method to stabilize such ill-conditioned calculations is closely related to Tikhonov stabilization (Tikhonov & Arsenin, 1977), where additional, *a priori* knowledge about the expected magnitude of the physical effect modelled is used to restrain the solutions, based on the same argument as in the case of restraints in the atomic refinement.

For example, logarithms of scale factors typically do not fluctuate by more than w_s between frames, where expectation about w_s is a function of the data-collection stability (beam stability, goniostat and/or crystal vibrations). This knowledge is described by adding a penalty term (*scale restrain*) to the functions being optimized:

$$(1/w_s^2)(p_{i_s} - p_{(i+1)_s})^2. \tag{11.4.8.9}$$

A similar approach can be used in calculations of absorption coefficients. For smooth absorption with the expectation of decreasing magnitude of parameters for high orders of spherical harmonics [equation (11.4.8.7)], a reasonable restraint term parameterized by w_a results in

$$\frac{l^2(p^2_{as,lm} + p^2_{ac,lm})}{w_a^2}. \tag{11.4.8.10}$$

If we do not want to penalize high-order terms more than the low-order ones, the following restraint can be used:

$$\frac{(p^2_{as,lm} + p^2_{ac,lm})}{w_a^2}. \tag{11.4.8.11}$$

11.4.9. Global refinement or post refinement

The process of refining crystal parameters using the combined reflection intensity measurements is known as global refinement or post refinement (Rossmann, 1979; Evans, 1993). The implementation of this method in *SCALEPACK* allows for separate refinement of the orientation of each image, but with the same unit-cell value for the whole data set. In each batch of data (a batch is typically one image), different unit-cell parameters may be poorly determined. However, in a typical data set there are enough orientations to determine all unit-cell lengths and angles precisely. Global refinement is also more precise than the processing of a single image in the determination of crystal mosaicity and the orientation of each image.

11.4.10. Merging – assessment of the error model and signal magnitudes in the data

Proper error estimation requires the use of Bayesian reasoning and a multi-component error model (Schwarzenbach *et al.*, 1989; Evans, 1993). In principle, the error estimates may be derived solely from a theoretical understanding of the measurement process. However, the complexity of error propagation and correlations between various sources of effects have led crystallographers to rely on hybrid approaches also involving self-consistency analysis of symmetry-equivalent reflections.

11.4.10.1. Estimation of random errors

The random errors in *DENZO* are estimated by a heuristic procedure that also accounts for small components of systematic errors (Borek *et al.*, 2003). Initially, *DENZO* estimated errors of integrated diffraction peaks from X-ray film. After introducing detectors with larger dynamic range, the procedure was adjusted accordingly.

The initial estimates of errors are obtained by

$$\sigma_0 = \frac{1}{\sum_i P_i^2/(B_i + P_i I)} \times \left\{ e_d \left[\sum_i P_i^2 (B_i + P_i I) + \frac{e_d}{n_b} \sum_i \frac{P_i^2 B_i}{(B_i + P_i I)^2} \right] \right\}^{1/2}, \tag{11.4.10.1}$$

where n_b is the number of pixels used in background estimation and e_d is the error-density parameter defined for each instrument, which can also be overridden by the user (Gewirth, 2003) with other variables defined in equation (11.4.7.1). The sums are calculated over all the pixels in a reflection profile. The expression within the braces { } describes two components of uncertainty: the left sum accounts for contributions resulting from pixels in the peak area, whereas the right sum adds an adjustment resulting from uncertainty of the background estimate. The denominator in the front of the expression in braces is derived from error propagation for the profile-fitted intensity [equation (11.4.7.3)].

Next, the goodness-of-profile-fitting factor g is calculated:

$$g = \left[\frac{1}{(n_i - 1)} \sum_i \frac{(M_i - B_i - P_i I)^2}{e_d (B_i + P_i I)} \right]^{1/2}, \tag{11.4.10.2}$$

where n_i is the number of pixels in a reflection profile. For weak reflections the parameter g should be relatively close to 1. If it is systematically off by a large factor, the error-density parameter e_d should be adjusted (Borek *et al.*, 2003). *SCALEPACK* applies an additional level of adjustment to the estimates produced by *DENZO* (Borek *et al.*, 2003):

$$\sigma_S = 1.2 \sigma_o g^{1/2}, \tag{11.4.10.3}$$

which is scaled either by the user or by an automatically adjustable factor E_S (called the error scale factor) to make disagreements among symmetry-related measurements consistent:

$$\sigma_I = E_S \sigma_S. \tag{11.4.10.4}$$

Even this scaled estimate of random error σ_I does not account for all types of errors and additional adjustments for systematic effects are needed.

11.4.10.2. Estimation of multiplicative errors

The multiplicative scale factor has its own uncertainty independent of random errors with typical values in the range of a few per cent. However, even such small errors are important in calculations of the phase signal. Errors in the scale factors have a correlated component that equally affects measurements of intensities in phasing differences, so it does not impact on the differences themselves. The important part is estimating the magnitude of the remaining component of scaling errors, described by σ_K. Comparing symmetry-related reflections estimates only the relevant component of multiplicative errors. The total scaling error would have to be estimated differently, but typically it has little relevance to macromolecular crystallography and can be ignored.

The σ_I [equation (11.4.10.4)] can be combined with σ_K to obtain the final estimated error of the scaled measurement:

$$\sigma_E = (1/K)(\sigma_I^2 + I^2 \sigma_K^2)^{1/2}. \tag{11.4.10.5}$$

11.4.10.3. Merging and signal validation

Symmetry-related scaled measurements $I(hkl)$ and their uncertainty estimates σ_E are used to obtain merged intensities by a standard weighted averaging formula:

$$\langle I \rangle = \frac{\sum_j I_j/\sigma_{E,j}^2}{\sum_j 1/\sigma_{E,j}^2}. \qquad (11.4.10.6)$$

This allows for calculations of validation statistics, called goodness-of-fit or normalized χ^2, for each unique index:

$$\chi^2 = \left(\frac{1}{n-1}\right)^{1/2} \sum_j \frac{(I_j - \langle I \rangle)^2}{\sigma_{E,j}^2}, \qquad (11.4.10.7)$$

where n represents the number of observations of a given unique index. This χ^2 statistic is then averaged in resolution shells or over intensity bins or batch number. If the error model accounts properly for all effects, the χ^2 statistic should fluctuate around a value of unity. If χ^2 values depart from this expectation it may indicate a number of possibilities, *e.g.* various problems at earlier stages (poorly edited beam-stop shadow, hardware failures, mistakes in processing or other source of outliers *etc.*), inadequacy of the error model or variations in the structure factors within the symmetry-related observations. The instrumental problems or mistakes in processing should be corrected. The effects that cannot be corrected may be handled by adjusting the error model. However, if the more detailed analysis eliminates the obvious source for such problems, then the most likely source of discrepancies between symmetry-related measurements results from violation of Friedel symmetry. *SCALEPACK* calculates merging statistics both for the Bijvoet pairs merged together and separately. Differences in χ^2 values between these two merging outputs are very reliable estimates of anomalous signal strength. When a more detailed analysis eliminates the obvious reasons for high χ^2 values, the most likely remaining source of error is non-isomorphism (Borek *et al.*, 2007, 2010).

11.4.11. Detector diagnostics

The *HKL* package has a number of tools that can detect possible detector or experimental setup problems (Otwinowski & Minor, 1997; Otwinowski *et al.*, 2003). Visual inspection of the image may provide only a very rough estimate of data quality. A check of the analogue-to-digital converter can provide rough diagnostics of detector electronics. Examination of the background can provide information about detector noise, especially when uncorrected images can be examined in the areas exposed to X-rays and areas where pure read-out noise can be observed. *DENZO* provides several diagnostic tools during the integration stage, as the crystallographer may observe crystal slippage, a change of unit-cell parameters or a change of the values of positional and angular χ^2 during the refinement.

Even more tools are provided at the data-scaling stage. By observing scale factors, poor crystal alignment can be detected. Other tools may help diagnose X-ray shutter malfunction, spindle-axis alignment and internal detector-alignment problems. The final inspection of outliers may again provide valuable information about detector quality. The clustering of outliers in one area of the detector may indicate a damaged surface; if most outliers are partials, it may indicate a problem with spindle backlash or shutter control. The zoom mode may be used to display the area around the outliers to identify the source of a problem: for example, the existence of a satellite crystal or single pixel spikes due to electronic failure. Sometimes, even for very strong data, a histogram of the pixel intensities may stop below the maximum valid pixel value, indicating saturation of the data-acquisition hardware or software.

11.4.12. *HKL-2000* and *HKL-3000*

DENZO and *SCALEPACK* form the numerical processing core of the *HKL* package. These programs can be used directly by editing commands in input scripts, but most of the time they are run through *HKL-2000* and its more expanded version *HKL-3000*.

The basic mode of *HKL-2000* reads in previously collected data as input and produces scaled and merged reflections as output. This use can be extended in three directions: control of the data-collection process by *HKL-2000* and *HKL-3000* (Minor *et al.*, 2002, 2006), incorporation of methods of structure solution by *HKL-3000* (Minor *et al.*, 2006), and storage of critical intermediate results in an external database (Grabowski *et al.*, 2007).

Crystallographic structure determinations encompass a wide range of project dynamics. There are many projects where all the steps are executed serially, while others involve substantial iterative improvements, where one or a few stages are repeated a number of times. *HKL-2000* and *HKL-3000* are designed to make both types of project more effective. To accomplish this, information is automatically propagated between various stages of analysis, and many necessary data transformations are performed to accommodate the interface requirements of many programs and beamline controls. At the same time, crystallography requires the experimenter to be actively engaged in decision making, depending on the nature of a particular project and types of problems encountered. The experimenter may need to assess the quality of a set of crystals, decide how to collect data sets from a chosen subset, determine the symmetry of each diffraction pattern, reassess the crystal quality based on the integration and merging steps, and then solve the structure using an appropriate method. Not all programs are fully automatic, and the experimenter may need to be involved in defining how a particular procedure should be executed. As a consequence, both *HKL-2000* and *HKL-3000* have a multiplicity of interface screens (accessed by tabs in the graphical control centre), each of them designed to control a particular process.

The versions of *HKL-2000* and *HKL-3000* that interface with data-collection systems can coordinate all parameters of the diffraction experiment. This facilitates interactive experiments in which data analysis is done online and results are automatically updated when new data are collected. In such experiments, it is possible to adjust the data-collection strategy to guarantee the desired result, particularly with regard to data completeness. The strategy takes into account limitations arising from radiation damage. Radiation damage can be estimated from past experience with similar crystals, by theoretical calculations of decay based on beam intensity, and by evaluating scale- and B-factor changes in real time.

The graphical control centre of *HKL-2000* (and *HKL-3000*) consists of three components: an internal database (optionally connected to an external one), a transition-state engine and a graphical user interface (GUI). The internal database stores all the information about data processing and data collection. It can describe not only the data already collected, but also those being collected and even those planned for collection. Each datum entered or program executed, including the data-collection interface, induces a change in the database by the transition-state engine. One of the main functions of the GUI is input to and

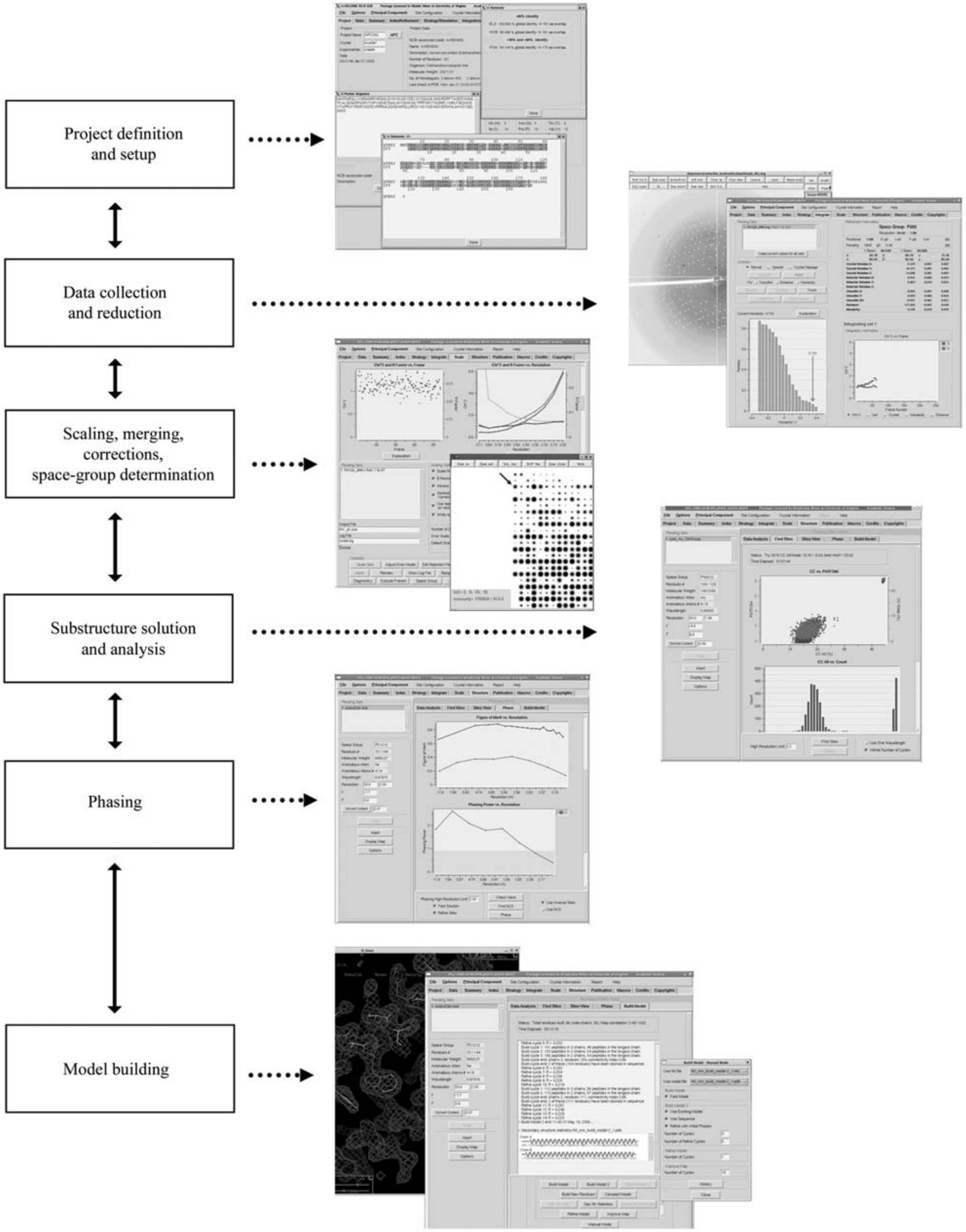

Project definition and setup
Data collection and reduction
Scaling, merging, corrections, space-group determination
Substructure solution and analysis
Phasing
Model building

Figure 11.4.12.1
General layout of *HKL-3000*. The experimenter is not limited to movement between neighbouring steps. For example, using feedback from model building, one can decide to collect more data (for example, a second wavelength or a low-resolution pass) and jump directly back to the 'Data collection and reduction' step.

editing of the database. The other major function is generation of reports from the database (to visualize its status).

The internal database abstraction is based on the following hierarchy: instrument type; site; experiment; crystal; three-dimensional (3D) group of diffraction images; and diffraction image. Each lower level of the hierarchy inherits the properties of the higher levels. When a program finishes analysing data at a particular level, the parent (higher-level) data are updated, so that data at the same level communicate only through the change of state of their common parent. The site-level data are created only when diffraction from a new detector is seen or when the parameters of the detector are changed, which is done rarely and typically by the X-ray equipment administrator. The experiment-level data describe diffraction data from one or more crystals of the same space group. A uniform series of diffraction images form 3D groups. There is no limit to the number of 3D groups, and, in the case of non-uniformity in the series (*e.g.* as found during data analysis), one 3D group can be split into two or more smaller 3D groups. The smallest 3D group can consist of one image. A crystal datum contains a set of 3D groups with a relative orientation and exposure level known *a priori*. In practice, this means that diffraction data encapsulated within a single crystal datum were collected from one sample at one site with potentially different settings of goniostat, data-collection axis, crystal translation, detector position, detector mode (*e.g.* binned/unbinned) or exposure level.

HKL-3000 combines a number of existing macromolecular crystallographic computer programs [*SHELX* (Sheldrick, 2008), *CCP4* (Collaborative Computational Project, No. 4, 1994), *SOLVE/RESOLVE* (Terwilliger, 2004), *ARP/wARP* (Perrakis *et al.*, 1999) and *COOT* (Emsley & Cowtan, 2004)] and decision-making algorithms into a powerful expert system (Fig. 11.4.12.1). The typical end result of *HKL-3000* is an interpretable electron-density map with a partially built structure and, in some cases, an almost complete and refined model.

The *HKL-3000* system is designed to evaluate the results of laboratory and synchrotron diffraction experiments very quickly. The system supports most common types of macromolecular crystallography experiments: isomorphous (IR) and molecular replacement (MR), multiple anomalous diffraction (MAD)/single anomalous diffraction (SAD), and native data collection for high-resolution refinement of previously solved models of proteins or protein–ligand complexes. SAD experiments have became popular in recent years (Chruszcz *et al.*, 2008) at least partly due to development of integrated systems like *HKL-3000*, which allow the experimenter to validate whether experimental data collected at one wavelength are of sufficient quality to solve the structure. The concurrent data collection, data processing and quick preliminary structure solution made by *HKL-3000* verifies the success of the X-ray experiment and allows optimization of the data-collection strategy while the crystal is still on the goniostat. This allows the experimenter to decide whether the experiment has been successfully completed and thus whether the crystal can be removed.

11.4.13. Final note

The methods presented here have been used to solve tens of thousands of crystal structures deposited in public data repositories. The results range from crystal structures of inorganic molecules with 3 Å unit-cell parameters to a crystal structure of a 700 Å-diameter virus, which crystallized in a 700 × 1000 × 1400 Å unit cell. The precision of signal estimation was sufficient to determine a structure where the amplitudes of the imaginary Bijvoet components were approximately 0.15% of the total amplitudes (Borek *et al.*, 2007), and for the results of integration and scaling analyses to be used in charge-density studies (Dominiak *et al.*, 2006). The success of these methods is a consequence of the robust and stable implementation described above.

This work was supported by NIH grant GM-53163. We would like to acknowledge the contributions of the many researchers who provided us with diffraction data representing a wide range of experimental problems. We are indebted to the following people who developed other diffraction-data-analysis programs, interactions with whom contributed to *HKL* program development and to the ideas presented here: M. G. Rossmann, G. Bricogne, P. Evans, A. Howard, W. Kabsch, A. Leslie, J. Pflugrath and G. Sheldrick. We would also like to thank H. Czarnocka, R. Henderson, W. Majewski, A. Pertsemlidis, D. Tomchick and M. Zimmerman for help in preparing this manuscript.

References

Azaroff, L. V. (1955). *Polarization correction for crystal-monochromatized X-radiation. Acta Cryst.* **8**, 701–704.

Blessing, R. H. (1995). *An empirical correction for absorption anisotropy. Acta Cryst.* A**51**, 33–38.

Blum, M., Metcalf, P., Harrison, S. C. & Wiley, D. C. (1987). *A system for collection and on-line integration of X-ray diffraction data from a multiwire area detector. J. Appl. Cryst.* **20**, 235–242.

Borek, D., Cymborowski, M., Machius, M., Minor, W. & Otwinowski, Z. (2010). *Diffraction data analysis in the presence of radiation damage. Acta Cryst.* D**66**, 426–436.

Borek, D., Ginell, S. L., Cymborowski, M., Minor, W. & Otwinowski, Z. (2007). *The many faces of radiation-induced changes. J. Synchrotron Rad.* **14**, 24–33.

Borek, D., Minor, W. & Otwinowski, Z. (2003). *Measurement errors and their consequences in protein crystallography. Acta Cryst.* D**59**, 2031–2038.

Bricogne, G. (1987). *The EEC Cooperative Programming Workshop on Position-sensitive Detector Software.* In *Proceedings of the Daresbury Study Weekend at Daresbury Laboratory*, 23–24 January, edited by J. R. Helliwell, P. A. Machin & M. Z. Papiz, pp. 120–146. Warrington: Daresbury Laboratory.

Chruszcz, M., Wlodawer, A. & Minor, W. (2008). *Determination of protein structures – a series of fortunate events. Biophys. J.* **95**, 1–9.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

Dominiak, P. M., Makal, A., Mallinson, P. R., Trzcinska, K., Eilmes, J., Grech, E., Chruszcz, M., Minor, W. & Wozniak, K. (2006). *Continua of interactions between pairs of atoms in molecular crystals. Chem. Eur. J.* **12**, 1941–1949.

EEC Cooperative Workshop on Position-Sensitive Detector Software (1986). Phase I and II, LURE, Paris, 16 May – 7 June; Phase III, LURE, Paris, 12–19 November.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics. Acta Cryst.* D**60**, 2126–2132.

Evans, P. R. (1987). *Postrefinement of oscillation camera data.* In *Proceedings of the Daresbury Study Weekend at Daresbury Laboratory*, 23–24 January, edited by J. R. Helliwell, P. A. Machin & M. Z. Papiz, pp. 58–66. Warrington: Daresbury Laboratory.

Evans, P. R. (1993). *Data reduction: data collection and processing.* In *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, 29–30 January, edited by L. Sawyer, N. Isaac & S. Bailey, pp. 114–123. Warrington: Daresbury Laboratory.

Gewirth, D. (2003). *HKL Manual.* 6th ed. HKL Research, Charlottesville, USA.

Grabowski, M., Joachimiak, A., Otwinowski, Z. & Minor, W. (2007). *Structural genomics: keeping up with expanding knowledge of the protein universe. Curr. Opin. Struct. Biol.* **17**, 347–353.

Greenhough, T. J. (1987). *Partials and partiality.* In *Proceedings of the Daresbury Study Weekend at Daresbury Laboratory*, 23–24 January, edited by J. R. Helliwell, P. A. Machin & M. Z. Papiz, pp. 51–57. Warrington: Daresbury Laboratory.

Higashi, T. (1990). *Auto-indexing of oscillation images. J. Appl. Cryst.* **23**, 253–257.

Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H. & Salemme, F. R. (1987). *The use of an imaging proportional counter in macromolecular crystallography. J. Appl. Cryst.* **20**, 383–387.

Howard, A. J., Nielsen, C. & Xuong, Ng. H. (1985). *Software for a diffractometer with multiwire area detector. Methods Enzymol.* **114**, 452–472.

International Tables for Crystallography (2005). Vol. A. *Space-Group Symmetry*, edited by Th. Hahn. Heidelberg: Springer.

Kabsch, W. (1988). *Evaluation of single-crystal X-ray diffraction data from a position sensitive detector. J. Appl. Cryst.* **21**, 916–924.

Kabsch, W. (1993). *Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants. J. Appl. Cryst.* **26**, 795–800.

Katayama, C. (1986). *An analytical function for absorption correction. Acta Cryst.* A**42**, 19–23.

Kim, S. (1989). *Auto-indexing oscillation photographs. J. Appl. Cryst.* **22**, 53–60.

Kopfmann, G. & Huber, R. (1968). *A method of absorption correction by X-ray intensity measurements. Acta Cryst.* A**24**, 348–351.

Leslie, A. (1993). *Autoindexing of rotation diffraction images and parameter refinement.* In *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, 29–30 January, edited by L. Sawyer, N. Isaac & S. Bailey, pp. 44–51. Warrington: Daresbury Laboratory.

Leslie, A. G. W. & Tsukihara, T. (1980). *A strategy for collecting isomorphous derivative data with the oscillation method. J. Appl. Cryst.* **13**, 304–305.

Matthews, B. W. & Czerwinski, E. W. (1975). *Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta Cryst.* A**31**, 480–487.

Messerschmidt, A. & Pflugrath, J. W. (1987). *Crystal orientation and X-ray pattern prediction routines for area-detector diffraction systems in macromolecular crystallography. J. Appl. Cryst.* **20**, 306–315.

Minor, W., Cymborowski, M. & Otwinowski, Z. (2002). *Automatic system for crystallographic data collection and analysis. Acta Phys. Pol. A*, **101**, 613–619.

Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes. Acta Cryst.* D**62**, 859–866.

Naday, I., Ross, S., Westbrook, E. M. & Zentai, G. (1998). *Charge-coupled device/fiber optic taper array X-ray detector for protein crystallography. Opt. Eng.* **37**, 1235–1244.

Otwinowski, Z. (1993). *Oscillation data reduction program.* In *Proceedings of the Daresbury CCP4 Study Weekend. Data Reduction and Processing*, edited by L. Sawyer, N. Isaacs and S. Bailey, pp. 56–62. Warrington: Daresbury Laboratory.

Otwinowski, Z., Borek, D., Majewski, W. & Minor, W. (2003). *Multiparametric scaling of diffraction intensities. Acta Cryst.* A**59**, 228–234.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode. Methods Enzymol.* **276**, 307–326.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Automated protein model building combined with iterative structure refinement. Nat. Struct. Biol.* **6**, 458–463.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1989). *Numerical Recipes – the Art of Scientific Computing.* Cambridge University Press.

Reeke, G. N. Jr (1984). *Eigenvalue filtering in the refinement of crystal and orientation parameters for oscillation photography. J. Appl. Cryst.* **17**, 238–243.

Rossmann, M. G. (1979). *Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting. J. Appl. Cryst.* **12**, 225–238.

Sakabe, N. (1991). *X-ray diffraction data collection system for modern protein crystallography with a Weissenberg camera and an imaging plate using synchrotron radiation. Nucl. Instrum. Methods A*, **303**, 448–463.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Statistical descriptors in crystallography. Report of the IUCr Subcommittee on Statistical Descriptors. Acta Cryst.* A**45**, 63–75.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Steller, I., Bolotovsky, R. & Rossmann, M. G. (1997). *An algorithm for automatic indexing of oscillation images using Fourier analysis. J. Appl. Cryst.* **30**, 1036–1040.

Terwilliger, T. (2004). *SOLVE and RESOLVE: automated structure solution, density modification and model building. J. Synchrotron Rad.* **11**, 49–52.

Tikhonov, A. N. & Arsenin, V. I. A. (1977). *Solutions of ill-posed problems.* Washington, New York, Winston: Halsted Press.

Vriend, G. & Rossmann, M. G. (1987). *Determination of the orientation of a randomly placed crystal from a single oscillation photograph. J. Appl. Cryst.* **20**, 338–343.

Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *The oscillation method for crystals with very large unit cells. Acta Cryst.* A**35**, 901–911.

International Tables for Crystallography (2012). Vol. F, Chapter 11.5, pp. 296–303.

Chapter 11.5. The use of partially recorded reflections for post refinement, scaling and averaging X-ray diffraction data

C. G. van Beek, R. Bolotovsky and M. G. Rossmann

11.5.1. Introduction

Recent advances in the use of frozen crystals of biological samples for X-ray diffraction data collection (Rodgers, 1994) often result in data for which most of the observed reflections on each frame are partially observed. This might be avoided by increasing the oscillation ranges, but this would cause many reflections to overlap with their neighbours. Hence, it is necessary to develop scaling procedures that are independent of the exclusive use of fully recorded reflections.

A set of measured Bragg intensities is dependent on the properties of the crystal, radiation source and detector. Usually, these factors cannot be kept constant throughout the data collection. The crystal may decay, weakening the Bragg intensities, or even 'die', which requires the use of several crystals for a full data set. The intensity and position of the primary X-ray beam may vary, especially at synchrotron-radiation sources. Finally, the detector response may change when, for example, different films or imaging plates are used during the data collection.

Most data sets can be divided into series of subsets, or frames, collected under more-or-less constant conditions. These frames need to be placed on a common arbitrary scale. The scaling can be performed by comparing the intensities of multiply measured reflections or symmetry-equivalent reflections on different frames.

A least-squares procedure frequently used for scaling frames of data is the Hamilton, Rollett and Sparks (HRS) method (Hamilton *et al.*, 1965). The target for the HRS least-squares minimization is

$$\psi = \sum_h \sum_i W_{hi}(I_{hi} - G_m I_h)^2, \qquad (11.5.1.1)$$

where I_h is the best estimate of the intensity of a reflection with reduced Miller indices h, I_{hi} is the intensity of the ith measurement of reflection h, W_{hi} is a weight for reflection h_i and G_m is the inverse linear scale factor for frame m on which reflection h_i is recorded. The reduced Miller indices are those corresponding to an arbitrarily defined asymmetric unit of reciprocal space. The HRS expression (11.5.1.1) assumes that all reflections h_i are full, that is, their reciprocal-lattice points have completely passed through the Ewald sphere.

For all unique reflections h, the values of I_h must correspond to a minimum in ψ. Thus,

$$\partial\psi/\partial I_h = 0. \qquad (11.5.1.2)$$

Therefore, the best least-squares estimate of the intensity of a reflection is

$$I_h = \sum_i W_{hi} G_m I_{hi} \Big/ \sum_i W_{hi} G_m^2. \qquad (11.5.1.3)$$

Since ψ is not linear with respect to the scale factors G_m, the values of the scale factors have to be determined by an iterative nonlinear least-squares procedure. As the scale factors are relative to each other, the HRS procedure requires that one of them be fixed.

Fox & Holmes (1966) describe an improved method of solving the HRS normal equations. Their approach is based on the singular value decomposition of the normal equations matrix. The advantage of the Fox and Holmes method, apart from the accelerated convergence of the least-squares procedure, is that no *ad hoc* decision needs to be made as to which scale factor should be fixed. Furthermore, 'troublesome' frames of data can be identified as causing negligibly small eigenvalues in the normal equations matrix.

11.5.2. Generalization of the Hamilton, Rollett and Sparks equations to take into account partial reflections

When a Bragg reflection is completely exposed within the oscillation range of one frame, a so-called 'full reflection', it gives rise to the 'full intensity'. In general, a Bragg reflection will occur on a number of consecutive frames as a series of partial reflections, and the full intensity can only be estimated from the measured intensities of the partial reflections. Let I_{him} represent the intensity contribution of reflection h_i recorded on frame m; if all the parts of h_i are available in the data set, then

$$I_{hi} = \sum_m (I_{him}/G_m). \qquad (11.5.2.1)$$

In practice, there will always be reflections that do not have all their parts available. In such cases, the only way to estimate the full intensity of a reflection is to apply an estimated value of the partiality to the measured reflection intensities.

Various models have been proposed to calculate the reflection partiality. Here we use Rossmann's model (Rossmann, 1979; Rossmann *et al.*, 1979) with Greenhough & Helliwell's (1982) correction. This model treats partiality as a fraction of a spherical volume swept through the Ewald sphere. The coordinates of the reciprocal-lattice point are defined by the Miller indices of the reflection, the crystal orientation matrix and the rotation angle. The volume of the sphere around the reciprocal-lattice point accounts for crystal mosaicity and beam divergence. Alternative geometrical descriptions of a reciprocal-lattice point passing through the Ewald sphere have been given by Winkler *et al.* (1979) and Bolotovsky & Coppens (1997).

Provided the reflection partiality, p_{him}, is known, the full intensity is estimated by

$$I_{hi} = I_{him} \big/ p_{him} G_m. \qquad (11.5.2.2)$$

This expression can produce as many estimates of I_{hi} as there are parts of reflection h_i, while expression (11.5.2.1) produces only one estimate of I_{hi} when all parts of reflection h_i are recorded. Having defined the relationships between measured intensities of partial reflections and estimated full reflections by expressions (11.5.2.1) and (11.5.2.2), two methods of generalizing the HRS equations can be considered.

Method 1. If a reflection h_i occurs on a number of consecutive frames and all parts of I_{him} are available in the data set, the generalized HRS target equation takes the form

Copyright © 2012 International Union of Crystallography

$$\psi = \sum_h \sum_i \sum_m W_{him} \left\{ I_{him} - G_m \left[I_h - \sum_{m' \neq m} (I_{him'} / G_{m'}) \right] \right\}^2 . \tag{11.5.2.3}$$

Using expression (11.5.1.2), the best least-squares estimate of I_h will be

$$I_h = \frac{\sum_i \left[\sum_m (I_{him}/G_m) \right] \left(\sum_m W_{him} G_m^2 \right)}{\sum_i \sum_m W_{him} G_m^2} = \frac{\sum_i I_{hi} \sum_m W_{him} G_m^2}{\sum_i \sum_m W_{him} G_m^2} . \tag{11.5.2.4}$$

Method 2. If the theoretical partiality, p_{him}, of the partially recorded reflection h_{im} can be estimated, the generalized HRS target equation takes the form

$$\psi = \sum_h \sum_i \sum_m W_{him} (I_{him} - G_m p_{him} I_h)^2 \tag{11.5.2.5}$$

and, using expression (11.5.1.2), the best least-squares estimate of I_h will then be

$$I_h = \frac{\sum_i \sum_m W_{him} G_m p_{him} I_{him}}{\sum_i \sum_m W_{him} G_m^2 p_{him}^2} . \tag{11.5.2.6}$$

When all reflections in the data set are fully recorded, expressions (11.5.2.3) and (11.5.2.5) reduce to the 'classical' HRS expression (11.5.1.1), and expressions (11.5.2.4) and (11.5.2.6) reduce to expression (11.5.1.3).

The scale factor G_m can be generalized to incorporate crystal decay (Gewirth, 1996; Otwinowski & Minor, 1997):

$$G_{him} = G_m \exp\left\{ -2B_m \left[\sin(\theta_{hi})/\lambda \right]^2 \right\}, \tag{11.5.2.7}$$

where B_m is a parameter describing the crystal disorder while frame m was recorded, θ_{hi} is the Bragg angle of reflection h_i and λ is the X-ray wavelength.

Method 1 only allows the refinement of the scale factors while method 2 allows refinement of the scale factors, crystal mosaicity and orientation matrix, as the latter two factors contribute to the calculated partiality. Furthermore, method 2 is essential for scaling of data sets with low redundancy (*e.g.* data collected from low-symmetry crystals or data collected over small rotation ranges). When a reflection h_i spans more than one frame, but there are no other reflections with the same reduced Miller indices h in the data set, the contribution of any partial reflection h_{im} to expression (11.5.2.3) will be zero, as in this case I_h will be the same as I_{hi}. In contrast, in method 2 the reflection h_i can be used for scaling because the estimates of the full intensity I_{hi} are calculated independently from every frame spanned by reflection h_i.

11.5.3. Selection of reflections useful for scaling

Both scaling methods 1 and 2 may take into account any reflection intensity observation, regardless of whether it is a partially or fully recorded reflection. However, there are significant differences between the selection of reflections in the two methods. Method 1 requires that all parts of a reflection are available in order to incorporate the reflection into the generalized HRS target, expression (11.5.2.3). Thus, reflections that occur at the beginning or the end of the crystal orientation, or at gaps within the rotation range, must be rejected. Even when all parts of a reflection are recorded, there might be parts for which there was a problem during integration, thus making the reflection useless for scaling. The decision on whether all parts of a reflection are available for scaling is dependent on knowledge of the crystal mosaicity and of the crystal orientation matrix. Since these might be inaccurate, a reasonable tolerance has to be exercised when deciding if a reflection has been completely measured on consecutive frames. Method 2 allows the use of all reflections for scaling as every observation of a partial reflection is sufficient to estimate the intensity of a full reflection, expression (11.5.2.2). However, a reasonable lower limit of calculated partiality has to be imposed in selecting reflections useful for scaling. The criteria for rejecting reflections prior to scaling and averaging are listed in Table 11.5.3.1.

11.5.4. Restraints and constraints

Scale factors will depend on the variation of the incident X-ray beam intensity, crystal absorption and radiation damage. Hence, in general, scale factors can be constrained to follow an analytical function or restrained to minimize variation between successive frames. The scale factors can be restrained by adding a term $w(G_n - G_{n+1})^2$ to ψ, expression (11.5.1.1), where G_n and G_{n+1} are scale factors for the nth and $(n+1)$th frame and w is a suitably chosen weight. Such procedures will increase $R_{\rm merge}$ but will also increase the accuracy of the scaled intensities as additional reasonable physical conditions have been applied.

The mis-setting angles of a single crystal should remain constant throughout the data set. Thus, in principle, the mis-setting angles should be constrained to be the same for all frames associated with a single crystal in the data set. However, in practice, independent refinement of the mis-setting angles can detect problems in the data set when there are discontinuities in

Table 11.5.3.1

Hierarchy of criteria for selecting reflections for scaling and averaging procedures

<table>
<tr><td colspan="2">Methods 1 and 2</td></tr>
<tr><td colspan="2">All parts of a reflection are rejected if:
(1) There are no successfully integrated parts.
(2) There are no parts with significant intensity (for scaling only).
(3) There are some parts entering and some parts exiting the Ewald sphere (this implies that the reflection is too close to the rotation axis and is partly in the blind zone).
(4) This is a full reflection recorded only once with no other symmetry-equivalent observations.</td></tr>
<tr><td>Method 1</td><td>Method 2</td></tr>
<tr><td>All parts of a reflection are rejected if:
(1) There is a part that is not successfully integrated.
(2) There is a part that has a significant intensity, but is not predicted by the crystal orientation and mosaicity used in the scaling program.
(3) The sum of calculated partialities differs from unity by more than a chosen value.</td><td>Any part of a reflection is rejected if:
(1) The calculated partiality is less than a chosen value.
(2) The intensity is less than a chosen fraction of the error estimate.</td></tr>
</table>

these angles with respect to frame number. Cell dimensions should be the same for all crystals and might therefore be constrained. However, care should be taken, as the exact conditions of freezing may cause some variations in cell dimensions between crystals. As radiation damage proceeds, mosaicity is likely to increase. Hence, constraint between the refined mosaicities of neighbouring frames can be useful.

11.5.5. Generalization of the procedure for averaging reflection intensities

Once the scale factors of all frames are determined, they need to be applied to the reflection intensities and error estimates. The reflection intensities with the same reduced Miller indices can then be averaged.

When method 2 is used for averaging, the determination of $\langle I_h \rangle$ is more complicated as there are as many estimates of the full intensity I_{hi} as there are partial reflections h_{im}. Therefore, intensity averaging of reflection h has to be done in two steps. First, for every reflection h_i, the intensity estimates from all partial observations will be the weighted mean, where the weights are based on the estimated standard deviations of each intensity measurement. In the second step, the average is taken over the i different scaled intensities for the observed reflections.

The selection of reflections useful for averaging is the same as for scaling (Table 11.5.3.1), except that it is no longer necessary to reject reflections that have insignificant intensities. Applying a σ cutoff while averaging the scaled intensities will lead to a statistical bias of the weaker reflection intensities.

For samples of three or more equivalent reflections, it is necessary to consider the absolute values of the differences between individual intensities and the median of the sample: $|I_{hi} - I_{\rm median}|$. The outliers can be detected by several statistical tests and, once detected, can be either down-weighted or rejected. When the sample consists of only two reflections, they can be considered a ‘discordant pair’ if the difference between their intensities is not warranted by the estimated errors and, hence, both reflections can be rejected (Blessing, 1997).

Averaging intensities estimated according to method 2 has an advantage over method 1 as outliers and discordant pairs can be ‘screened’ at two levels: firstly, when the estimates of the full reflection intensity I_{hi}, calculated by expression (11.5.2.2) from different parts of the same reflection, are considered, and secondly when the mean intensities $\langle I_{hi} \rangle$ from different reflections are considered.

11.5.6. Estimating the quality of data scaling and averaging

A commonly used estimate of the quality of scaled and averaged Bragg reflection intensities is $R_{\rm merge}$. Useful definitions of R factors are:

$$R_{\rm merge} = \quad R_1 = \left[\left(\sum_h\sum_i |I_{hi} - \langle I_h \rangle|\right)\Big/\sum_h\sum_i |I_{hi}|\right] \times 100\%, \qquad (11.5.6.1)$$

$$R_2 = \left\{\left[\sum_h\sum_i (I_{hi} - \langle I_h \rangle)^2\right]\Big/\sum_h\sum_i I_{hi}^2\right\} \times 100\% \qquad (11.5.6.2)$$

and

$$R_w = \left\{\left[\sum_h\sum_i W_{hi}(I_{hi} - \langle I_h \rangle)^2\right]\Big/\sum_h\sum_i W_{hi} I_{hi}^2\right\} \times 100\%. \qquad (11.5.6.3)$$

The linear (R_1), square (R_2) and weighted (R_w) R factors can be subdivided into resolution ranges, intensity ranges, reflection classes, frame number and regions of the detector surface. When method 1 is used, reflections h_i can be grouped in terms of the sums of partialities of contributing partial reflections h_{im}.

The R-factor variation depends on the properties of the detector with respect to intensities. Generally the R factor decreases as intensity increases. Thus, the R factor generally increases with resolution. Any deviation from this behaviour might indicate a problem in the data collection due to nonlinearity of the detector response, ice diffuse diffraction, or any other stray effects superimposed on the crystal diffraction.

A useful indicator of the quality of the intensity estimates of partial reflections is the mean ratio of calculated partiality to observed partiality:

$$r_p = \langle p_{him}^{\rm calc}/p_{him}^{\rm obs} \rangle = \langle p_{him}^{\rm calc} \langle I_h \rangle / I_{him} \rangle. \qquad (11.5.6.4)$$

The deviation of this ratio from unity can be examined as a function of the reflection intensity, resolution and calculated partiality.

The comparison of R factors for centric and noncentric reflections can be used to determine the significance of an anomalous-scattering effect. The quality of the anomalous-dispersion signal can be assessed by calculation of the scatter, σ_{Ih}, where

$$\sigma_{Ih} = \left\{[1/(n-1)]\sum_n (\langle I_h \rangle - I_{hn})^2\right\}^{1/2} \qquad (11.5.6.5)$$

and $\langle I_h \rangle$ is the average of the n measurements of the full reflection intensities I_{hn}. The σ_{Ih} values for noncentric reflections can be compared to the scatter, σ_{Ih}^+ or σ_{Ih}^-, of reflections differing only in absorption while excluding Bijvoet opposites. The mean scatter is calculated from all σ_{Ih} values,

$$\langle \sigma_{Ih} \rangle = (1/h)\sum_h \left\{[1/(n-1)]\sum_n (\langle I_h \rangle - I_{hn})^2\right\}^{1/2}. \qquad (11.5.6.6)$$

The ratios $\langle \sigma_{Ih} \rangle / \langle \sigma_{Ih}^+ \rangle$ and $\langle \sigma_{Ih} \rangle / \langle \sigma_{Ih}^- \rangle$ should be larger than unity for significant anomalous-dispersion data.

11.5.7. Experimental results

11.5.7.1. Variation of scale factors versus frame number

If scale factors are to make physical sense, their behaviour with respect to the frame number has to be in accordance with the known changes in the beam intensity, crystal condition and detector response.

The scaling of a φX174 procapsid data set (Dokland *et al.*, 1997) was performed using methods 1 and 2 as described here and using *SCALEPACK* (Otwinowski & Minor, 1997) (Fig. 11.5.7.1). Graphs (*a*) and (*b*) in Fig. 11.5.7.1 have four segments corresponding to four synchrotron beam ‘fills’. All three methods give scale factors within 5% of each other (Figs. 11.5.7.1*c* and *d*). However, for the first and last frame of each ‘fill’ the results can differ by as much as 15%. Both method 1 and *SCALEPACK* produce physically wrong results in that the scale factors of these frames look like outliers compared to the scale factors of the neighbouring frames. By contrast, method 2 provides consistent scale factors for these frames. Although the algorithm used by *SCALEPACK* for scaling frames with partial reflections has never been disclosed, the similar results obtained by method 1 and *SCALEPACK* suggest that *SCALEPACK* might be using an algorithm similar to that of method 1 (Fig. 11.5.7.1*d*).

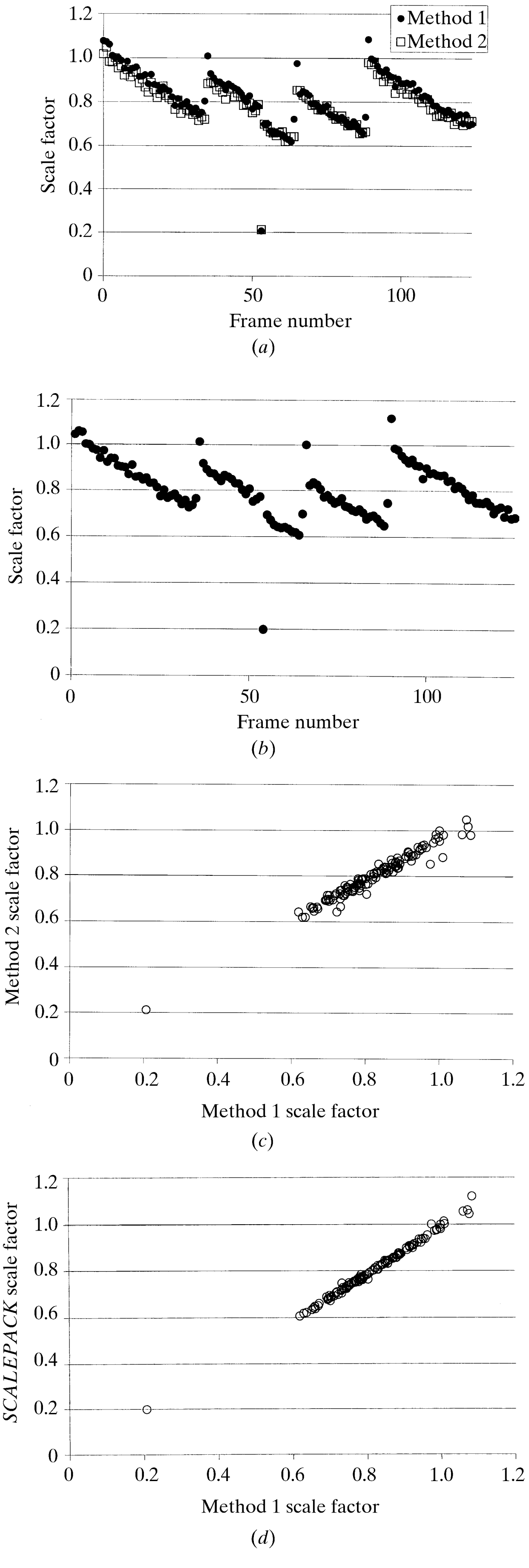


Figure 11.5.7.1

Linear scale factors as a function of frame number for a φX174 data set (Dokland *et al.*, 1997). Results from (*a*) method 1 and method 2, (*b*) *SCALEPACK*. Comparison of (*c*) method 2 *versus* method 1, and (*d*) *SCALEPACK versus* method 1.

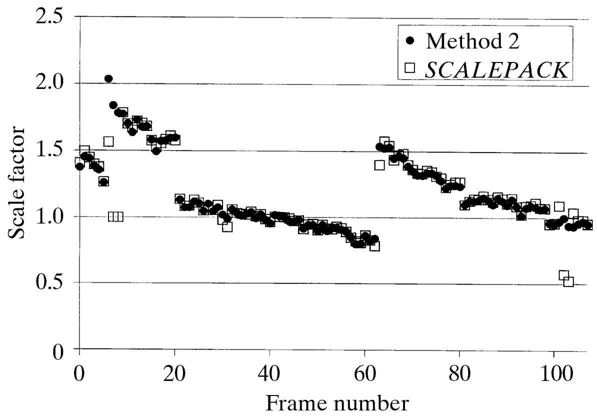


Figure 11.5.7.2

Linear scale factor as a function of frame number for an HRV14 data set (Rossmann *et al.*, 1985, 1997).

Attempts at scaling a data set of a frozen crystal of HRV14 (Rossmann *et al.*, 1985, 1997) failed with method 1 as a result of gaps in the rotation range for the first 20 frames, causing singularity of the normal equations matrix. When frames without useful neighbours were excluded, the cubic symmetry of the crystal was sufficient for successful scaling. In contrast, method 2 did not have any problems with the whole data set, and the results obtained with method 2 showed greater consistency than those obtained with method 1 or *SCALEPACK* (Fig. 11.5.7.2).

The accuracy and robustness of method 2 is also demonstrated by the scaling results for a Sindbis virus capsid protein (SCP), residues 114–264 (Choi *et al.*, 1991, 1996). The behaviour of the scale factor with respect to the frame number reflects the anisotropy of the thin plate-shaped crystal (Fig. 11.5.7.3). For the first 40 frames (frame numbers 0 to 39), even-numbered frames have higher scale factors than odd-numbered frames. Data collection was stopped after frame number 39 and restarted. After frame number 39, odd-numbered frames have higher scale factors than even-numbered frames. This effect presumably relates to the use of the two alternating image plates with slightly different sensitivities in the *R*-axis camera used in the data collection.

11.5.7.2. *R factor as a function of 'sum-of-partialities' (method 1)*

In order to determine the limits of tolerance that can be permitted when method 1 is used, the *R* factor was examined as a

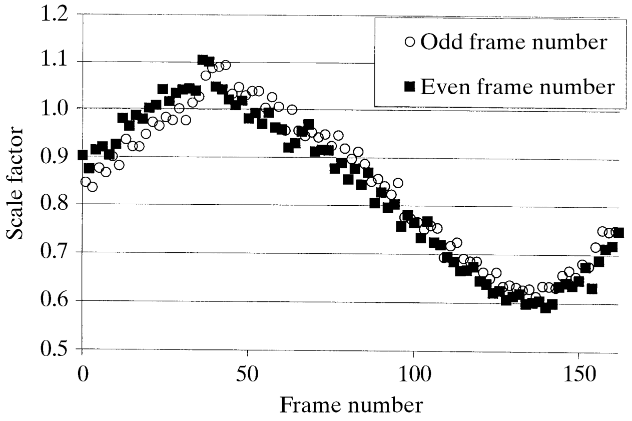


Figure 11.5.7.3

Linear scale factor determined by method 2 as a function of frame number for an SCP(114–264) data set (Choi *et al.*, 1991, 1996). The sine-like pattern reflects the anisotropy of a thin plate-shaped crystal.

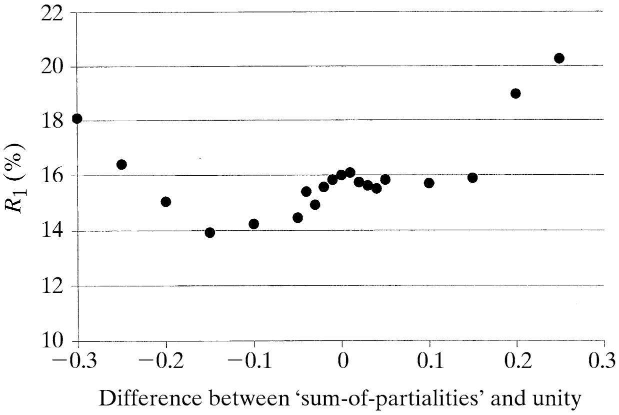


Figure 11.5.7.4
R factor as a function of the difference of calculated 'sum-of-partialities' and unity for the estimates of full reflections when method 1 is used for the scaling and averaging of a φX174 data set (Dokland *et al.*, 1997).

function of the sum-of-partialities for the φX174 procapsid data (Fig. 11.5.7.4). Reflections with sum-of-partialities of 1 ± 0.3 were used. The R factor changes sharply when the sum-of-partialities is outside 1 ± 0.15. Hence, ± 0.15 were acceptable limits of tolerance for this data set.

11.5.7.3. *Statistics for rejecting reflections and data quality as a function of frame number*

The behaviour of the R factor *versus* frame number (Fig. 11.5.7.5) is more monotonic when method 1 is used compared to method 2. In method 1, the data-quality estimates for neighbouring frames are strongly correlated because the full reflections used in the statistics are obtained by summing partials from consecutive frames. By contrast, in method 2 every frame produces estimates of full reflection intensities independently of the neighbouring frames. Therefore, the R factors per frame calculated after scaling with method 2 truly represent the data quality for individual frames.

11.5.7.4. *Observed versus calculated partiality*

The relationship between observed and calculated partialities (Fig. 11.5.7.6) deviates from the ideal line $p_{\rm obs} = p_{\rm calc}$, especially for the smaller calculated partialities where $p_{\rm obs} > p_{\rm calc}$. This suggests errors in the measurements of $p_{\rm obs}$ or the calculations of $p_{\rm calc}$. The latter may be improved by a post refinement of the orientation matrix and crystal mosaicity (Rossmann *et al.*, 1979).

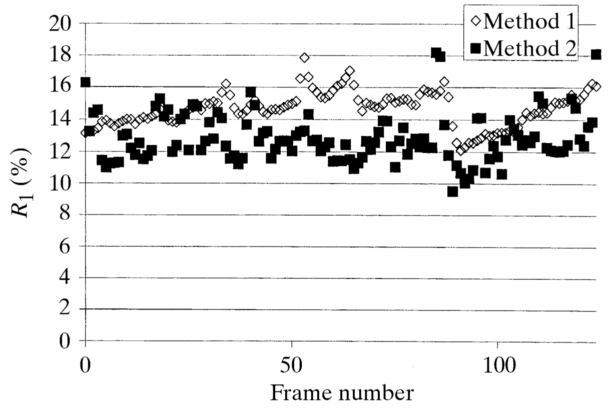


Figure 11.5.7.5
R factor per frame as a function of frame number for a φX174 data set (Dokland *et al.*, 1997).

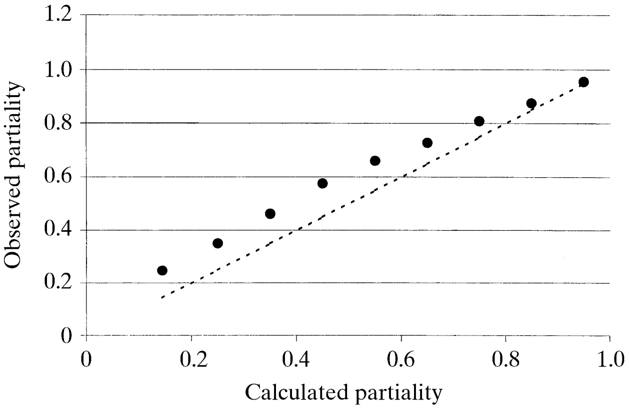


Figure 11.5.7.6
The observed partialities plotted against calculated partialities for a φX174 data set (Dokland *et al.*, 1997) processed by method 2. The observed partialities for individual partial reflections were averaged in bins of calculated partialities. The broken line represents the ideal relationship $p_{\rm obs} = p_{\rm calc}$.

11.5.7.5. *Anisotropic mosaicity*

Refinement of the effective mosaicity can show both the anisotropic nature of the crystal (Fig. 11.5.7.7) as well as the impact of radiation damage. The effective mosaicity is the convolution of the mosaic spread of the crystal, the beam divergence and the wavelength divergence of the incident X-ray beam. Hence, X-ray diffraction data collected at a synchrotron-radiation source necessitate the differentiation of the effective mosaicity in the horizontal and vertical planes. A more general approach is the introduction of six parameters reflecting the anisotropic effective mosaicity.

11.5.7.6. *Anomalous dispersion*

The quality of anomalous-dispersion data can be assessed by calculation of the average scatter, expression (11.5.6.6). The ratios $\langle \sigma_{Ih} \rangle / \langle \sigma^{+}_{Ih} \rangle$ and $\langle \sigma_{Ih} \rangle / \langle \sigma^{-}_{Ih} \rangle$ should be larger than unity for significant anomalous data (Fig. 11.5.7.8). Note the much larger ratios for the scatter among measurements of I_h for data measured at the absorption edge of Se, as opposed to measurements remote from the edge. The decreasing values of the ratios with resolution are due to the decrease of I_h value, thus causing

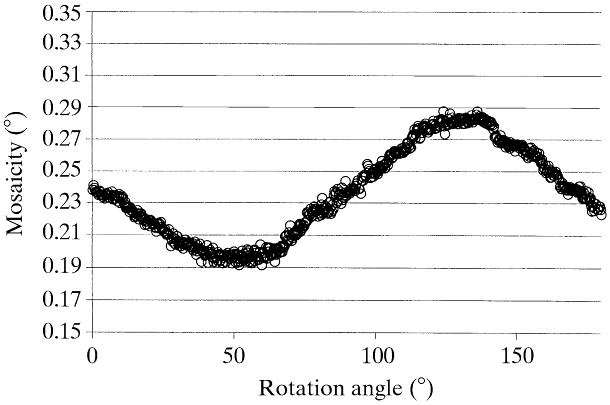


Figure 11.5.7.7
Variation of (unconstrained) mosaicity for a monoclinic crystal of the bacterial virus alpha3 (Bernal *et al.*, 1998) showing the crystal anisotropy.

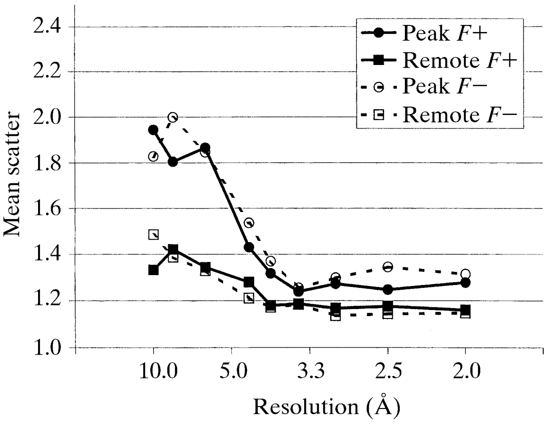


Figure 11.5.7.8
Quality of anomalous-dispersion data for an SeMet derivative of dioxygenase Rieske ferredoxin (Colbert & Bolin, 1999).

the error in the measurement of I_h to approach the difference in intensity of Bijvoet opposites.

11.5.8. Conclusions

The generalized HRS method allows scaling and averaging of X-ray diffraction data collected with an oscillation camera while simultaneously using full and partial reflections. The procedure is as useful for thin slices of reciprocal space as it is for thicker slices.

The results of data processing with the two different algorithms indicate that method 1, based on adding partial reflections, may fail to scale data sets with gaps in the rotation range or with low redundancy. The values of the scale factors obtained with both methods are similar, except for cases where there are gaps in the rotation range or dramatic changes in the true scale factors between consecutive frames. In these cases, method 1 produces a physically wrong result. The algorithm used by method 1 is probably similar to that used by *SCALEPACK* (Otwinowski & Minor, 1997).

Method 2 is more stable and versatile than method 1, and allows the scaling of data sets with incompletely measured reflections and low redundancy. The major drawback of method 2 is that errors in the crystal orientation matrix and mosaicity, as well as inadequacies of the theoretical model for reflection partiality, contribute to errors in the scaled intensities. Therefore, post refinement is needed for method 2 to perform at its best.

APPENDIX A11.5.1

Partiality model (Rossmann, 1979; Rossmann *et al.*, 1979)

Small differences in the orientation of domains within the crystal, as well as the cross fire of the incident X-ray beam, will give rise to a series of possible Ewald spheres. Their extreme positions will subtend an angle $2m$ at the origin of the reciprocal space, and their centres lie on a cusp of limiting radius $\delta = m/\lambda$, where m is the half-angle effective mosaic spread. As the reciprocal lattice is rotated around the axis (Oy) perpendicular to the mean direction of the incident radiation (Oz), a point P will gradually penetrate the effective thickness of the reflection sphere (Fig. A11.5.1.1). Initially, only a few domain blocks will satisfy Bragg's law, but upon further rotation the number of blocks that are in a reflecting condition will increase. The maximum will be reached when the point P has penetrated halfway through the sphere's effective thickness, after which there will be a decline of the crystal volume able to diffract.

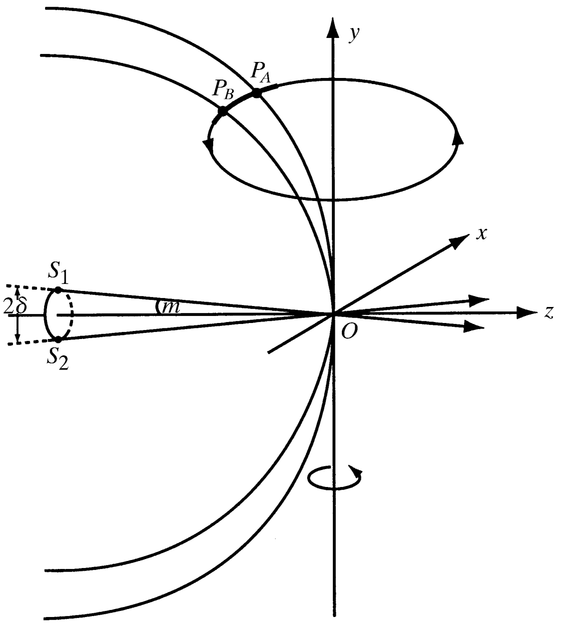


Figure A11.5.1.1
Penetration of a reciprocal-lattice point P into the sphere of reflection by rotation around Oy. The extremes of reflecting conditions at P_A and P_B are equivalent to X-rays passing along the lines S_1O and S_2O with centres of the Ewald spheres at S_1 and S_2 and subtending an angle of $2m$ at O. Hence, in three dimensions, the extreme reflecting spheres will lie with their centres on a circle of radius $\delta = m/\lambda$ at $z = -1/\lambda$.

Let q be a measure of the fraction of the path travelled by P between the extreme reflecting positions P_A and P_B, and let p be the fraction of the energy already diffracted. Then the relation between p and q must have the general form shown in Fig. A11.5.1.2. It is physically reasonable to assume that the curve for p is tangential to $q = 0$ at $p = 0$ and to $q = 1$ at $p = 1$.

A reasonable approximation to the above conditions can be obtained by considering the fraction of the volume of a sphere removed by a plane a distance q from its surface (Fig. A11.5.1.2). It is easily shown that if p is the volume, then

$$p = 3q^2 - 2q^3. \qquad \text{(A11.5.1.1)}$$

This curve is shown in Fig. A11.5.1.2 and corresponds to assuming that the reciprocal-lattice point is a sphere of finite volume cutting an infinitely thin Ewald sphere. Also shown in Fig. A11.5.1.2 is the line $p = q$ which would result if the reciprocal-lattice point were a rectangular block whose surfaces were parallel and perpendicular to the Ewald sphere at the point of penetration.

Assuming a right-handed coordinate system (x, y, z) in reciprocal space fixed to the camera, it is easily shown (Wonacott, 1977) that the condition for reflection is

$$d^{*2} + (2z/\lambda) = 0, \qquad \text{(A11.5.1.2)}$$

where d^* is the distance of a reciprocal-lattice point $P(x, y, z)$ from the origin, O, of reciprocal space. Similarly, it can be shown that at the ends of the path of the reciprocal-lattice point through the finite thickness of the sphere,

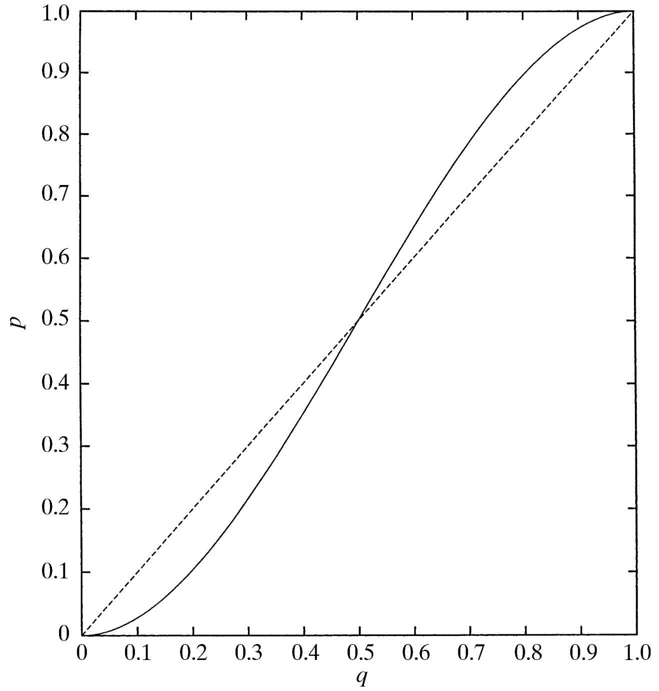


Figure A11.5.1.2
Relationship between the fraction of the path travelled, q, by a reciprocal-lattice point across an Ewald sphere of finite thickness and the fraction of the total scattered intensity, p. The curve shown is for $p = 3q^2 - 2q^3$. As an extreme case, the line $p = q$ is also shown.

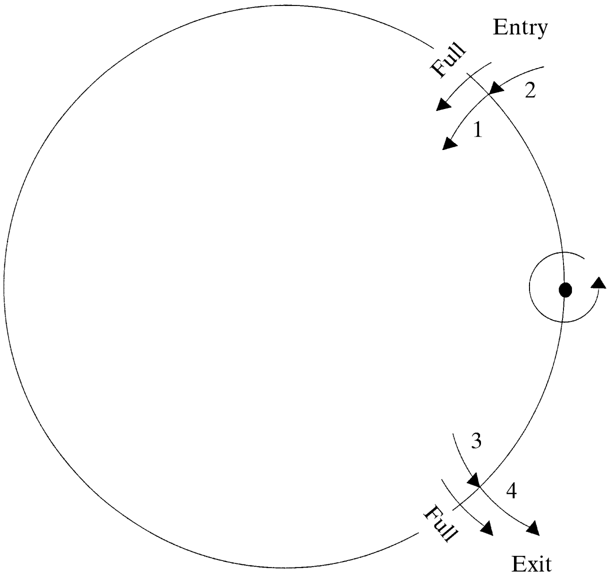


Figure A11.5.1.3
The four conditions 1, 2, 3 and 4 for partial reflections corresponding to Table A11.5.1.1. The arrow ends and heads correspond to the start and end positions of a reciprocal-lattice point, respectively.

$$\begin{aligned} &d^{*2} + \delta^2 + (2z/\lambda) - 2\delta\left(x_A^2 + y_A^2\right)^{1/2} = 0 \quad \text{and} \\ &d^{*2} + \delta^2 + (2z/\lambda) - 2\delta\left(x_B^2 + y_B^2\right)^{1/2} = 0. \end{aligned} \qquad \text{(A11.5.1.3)}$$

Therefore,

$$\begin{aligned} z_A &= (\lambda/2)\left[-d^{*2} - \delta^2 + 2\delta\left(x_A^2 + y_A^2\right)^{1/2}\right], \\ z_B &= (\lambda/2)\left[-d^{*2} - \delta^2 + 2\delta\left(x_B^2 + y_B^2\right)^{1/2}\right]. \end{aligned} \qquad \text{(A11.5.1.4)}$$

Since δ is small, it can be assumed that $2\delta(x^2 + y^2)^{1/2}$ is independent of the position of the reciprocal-lattice point P between the extreme positions P_A and P_B (Fig. A11.5.1.1). Hence, the length of the path through the finite thickness of the sphere is proportional to

$$z_A - z_B = 2\lambda\delta\left(x_P^2 - y_P^2\right)^{1/2}. \qquad \text{(A11.5.1.5)}$$

Now, if a reflection is only just penetrating the sphere at the end of the oscillation range, then the fraction of penetration is given by

$$q = PP_A/P_AP_B = (z_P - z_A)/(z_B - z_A). \qquad \text{(A11.5.1.6)}$$

Substituting this expression into equation (A11.5.1.4), it follows that

$$q = \tfrac{1}{2}\left[1 + (D_1/\eta_1)\right], \qquad \text{(A11.5.1.7)}$$

where

$$D = d^{*2} + \delta^2 + (2z/\lambda) \qquad \text{(A11.5.1.8)}$$

and

$$\eta = 2\delta(x^2 + y^2)^{1/2}. \qquad \text{(A11.5.1.9)}$$

The subscripts A and B refer to the beginning and end of the oscillation range for the partial reflection P, respectively.

Similarly, if a reflection is almost completely within the sphere,

$$q = PP_B/P_AP_B = (z_B - z_P)/(z_B - z_A) = \tfrac{1}{2}\left[1 - (D_2/\eta_2)\right]. \qquad \text{(A11.5.1.10)}$$

There are indeed four such conditions: two while a reflection is entering the Ewald sphere, and two while it is exiting. As such, it is readily seen that $-1 < (D_i/\eta_i) < 1$ ($i = 1$ or 2) is the range for a partial reflection. The full range of conditions is given in Table A11.5.1.1, as are the conditions for a full reflection.

This article is based primarily on the original publication by Bolotovsky *et al.* (1998). We are grateful for an NSF Grand Challenge grant in support of this work.

References

Bernal, R., Burch, A., Fane, B. & Rossmann, M. G. (1998). Unpublished results.

Blessing, R. H. (1997). *Outlier treatment in data merging. J. Appl. Cryst.* **30**, 421–426.

Table A11.5.1.1
Calculation of the degree of penetration of the Ewald sphere, q

The subscripts refer to the angles φ_1 and φ_2, designating the beginning and end of the oscillation range, respectively. See Fig. A11.5.1.3 for graphical representations of conditions 1 to 4.

	Almost completely within sphere	Almost completely outside sphere	Full reflection
Entering	Condition 1: $-1 < D_1/\eta_1 < +1$ and $D_2/\eta_2 \leq -1$	Condition 2: $-1 < D_2/\eta_2 < +1$ and $D_1/\eta_1 \geq +1$	$D_1/\eta_1 \geq +1$ and $D_2/\eta_2 \leq -1$
Exiting	Condition 3: $-1 < D_2/\eta_2 < +1$ and $D_1/\eta_1 \leq -1$	Condition 4: $-1 < D_1/\eta_1 < +1$ and $D_2/\eta_2 \geq +1$	$D_1/\eta_1 \leq -1$ and $D_2/\eta_2 \geq +1$

Bolotovsky, R. & Coppens, P. (1997). *The φ extent of the reflection range in the oscillation method according to the mosaicity-cap model*. *J. Appl. Cryst.* **30**, 65–70.

Bolotovsky, R., Steller, I. & Rossmann, M. G. (1998). *The use of partial reflections for scaling and averaging X-ray area dectector data*. *J. Appl. Cryst.* **31**, 708–717.

Choi, H. K., Lee, S., Zhang, Y. P., McKinney, B. R., Wengler, G., Rossmann, M. G. & Kuhn, R. J. (1996). *Structural analysis of Sindbis virus capsid mutants involving assembly and catalysis*. *J. Mol. Biol.* **262**, 151–167.

Choi, H. K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M. G. & Wengler, G. (1991). *Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion*. *Nature (London)*, **354**, 37–43.

Colbert, C. & Bolin, J. (1999). Unpublished results.

Dokland, T., McKenna, R., Ilag, L. L., Bowman, B. R., Incardona, N. L., Fane, B. A. & Rossmann, M. G. (1997). *Structure of a viral procapsid with molecular scaffolding*. *Nature (London)*, **389**, 308–313.

Fox, G. C. & Holmes, K. C. (1966). *An alternative method of solving the layer scaling equations of Hamilton, Rollett and Sparks*. *Acta Cryst.* **20**, 886–891.

Gewirth, D. (1996). *The HKL Manual. A Description of the Programs DENZO, XDSPLAYF and SCALEPACK*, 5th ed., pp. 87–90. Yale University, New Haven, USA.

Greenhough, T. J. & Helliwell, J. R. (1982). *Oscillation camera data processing: reflecting range and prediction of partiality. I. Conventional X-ray sources*. *J. Appl. Cryst.* **15**, 338–351.

Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *On the relative scaling of X-ray photographs*. *Acta Cryst.* **18**, 129–130.

Otwinowski, Z. & Minor, W. (1997). *Processing of X-ray diffraction data collected in oscillation mode*. *Methods Enzymol.* **276**, 307–326.

Rodgers, D. W. (1994). *Cryocrystallography*. *Structure*, **2**, 1135–1140.

Rossmann, M. G. (1979). *Processing oscillation diffraction data for very large unit cells with an automatic convolution technique and profile fitting*. *J. Appl. Cryst.* **12**, 225–238.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Structure of a human common cold virus and functional relationship to other picornaviruses*. *Nature (London)*, **317**, 145–153.

Rossmann, M. G., Leslie, A. G. W., Abdel-Meguid, S. S. & Tsukihara, T. (1979). *Processing and post-refinement of oscillation camera data*. *J. Appl. Cryst.* **12**, 570–581.

Rossmann, M. G., Momany, C. A., Cheng, B. & Chakravarty, S. (1997). Unpublished results.

Winkler, F. K., Schutt, C. E. & Harrison, S. C. (1979). *The oscillation method for crystals with very large unit cells*. *Acta Cryst.* A**35**, 901–911.

Wonacott, A. J. (1977). *Geometry of the oscillation method*. In *The Rotation Method in Crystallography*, edited by U. W. Arndt & A. J. Wonacott, pp. 75–103. Amsterdam: North Holland.

International Tables for Crystallography (2012). Vol. F, Chapter 11.6, pp. 304–310.

Chapter 11.6. *XDS*

W. Kabsch

11.6.1. Functional specification

The program package *XDS* (Kabsch, 1988*a*,*b*, 1993; Chapter 11.3) was developed for the reduction of single-crystal diffraction data recorded on a planar detector by the rotation method using monochromatic X-rays. It includes a set of three programs.

XDS accepts a sequence of adjacent non-overlapping rotation images from a variety of imaging-plate, charge-coupled-device (CCD), pixel and multiwire area detectors, infers crystal symmetry and metrics, and produces a list of corrected integrated intensities of the reflections occurring in the images in a nearly automatic way. The program assumes that each image covers the same positive amount of crystal rotation and that the rotation axis, incident beam and crystal intersect at one point, but otherwise imposes no limitations on the detector position, on the directions of the rotation axis and incident beam or on the oscillation range covered by each image.

XSCALE places the data sets obtained from processing with *XDS* on a common scale, optionally merges them into one or several sets of unique reflections and reports their completeness and the quality of the integrated intensities. It corrects the data for absorption effects, sensitivity variations in the detector plane and radiation damage. Optionally, it can correct reflections individually for radiation damage by extrapolation to their initial intensities at zero dose.

XDSCONV converts reflection data files as obtained from *XDS* or *XSCALE* into various formats required by software packages for crystal structure determination. It can generate test reflections or inherit previously selected ones which are used for the calculation of a free R factor to monitor the progress of structure refinement.

11.6.2. *XDS*

XDS is organized into eight steps (major subroutines) which are called in succession by the main program. Information is exchanged between the steps by files (see Table 11.6.2.1), which allows the repetition of selected steps with a different set of input parameters without rerunning the whole program. The files generated by *XDS* are either ASCII-type files that can be inspected and modified using a text editor, or binary control images saved as a byte-offset variant of the CBFlib format (Bernstein & Hammersley, 2005; Bernstein & Ellis, 2005). Such images are indicated by the file-name extension `.cbf` and can be looked at using the open-source program *XDS-Viewer* written by Michael Hoffer.

All files have a fixed name defined by *XDS*, which makes it mandatory to process each data set in a newly created directory in order to avoid name clashes. Clearly, one should not run more than one *XDS* job at a time in the same directory. Output files affected by rerunning selected steps (see Table 11.6.2.1) should also first be given another name if their original contents are meant to be saved.

Data processing begins by copying an appropriate input file into the new directory. Input-file templates are provided with the *XDS* package for a number of frequently used data-collection facilities. The copied input file must be renamed `XDS.INP` and edited to provide the correct parameter values for the actual data-collection experiment. All parameters in `XDS.INP` are named by keywords containing an equals sign as the last character and many of them will be mentioned here in context in order to clarify their meaning. Execution of *XDS* (JOB=XDS) invokes each of the eight program steps as described below. The results and diagnostics from each step are saved in files with the extension `.LP` attached to the program-step name. These files should always be studied carefully to see whether processing was satisfactory or, in the case of failure, to find out what could have gone wrong.

11.6.2.1. *XYCORR*

This program step calculates a lookup table of additive spatial corrections at each detector pixel which is stored in the files `X-CORRECTIONS.cbf`, `Y-CORRECTIONS.cbf`. Often, the data images have already been corrected for geometrical distortions, in which case *XYCORR* produces tables of zeros.

For spiral read-out imaging-plate detectors the small corrections resulting from radial (ROFF=) and tangential (TOFF=) offset errors of the scanner are computed.

For some multiwire and CCD detectors that deliver geometrically distorted images, corrections are derived from a calibration image (BRASS_PLATE_IMAGE=file name). This image displays the response to a brass plate containing a regular grid of holes which is mounted in front of the detector and illuminated by an X-ray point source. Clearly, the source must be placed exactly at the location to be occupied by the crystal during the actual data collection, as photons emanating from the calibration source are meant to simulate all possible diffracted beam directions. For visual control, spots that have been located and accepted from the brass-plate image by *XYCORR* are marked in the file `FRAME.cbf`.

The following problems can be encountered in this step.

(i) A misplaced calibration source can lead to an incorrect lookup table, impairing the correct prediction of the observed diffraction pattern in subsequent program steps.
(ii) An underexposed calibration image can result in an incomplete and unreliable list of calibration spots.

11.6.2.2. *INIT*

INIT determines three lookup tables, saved as the files `BLANK.cbf`, `GAIN.cbf` and `BKGINIT.cbf`, that are required by the subsequent processing steps for classifying pixels in the data images as background or belonging to a diffraction spot ('strong' pixels). These tables should be inspected with the *XDS-Viewer* program.

`BLANK.cbf` contains a lookup table of the detector noise. It is determined from a specific image recorded in the absence of X-rays (DARK_CURRENT_IMAGE=) or is assumed to be a constant derived from the mean recorded value in each corner of the data images.

Copyright © 2012 International Union of Crystallography

Table 11.6.2.1
Information exchange between program steps of *XDS*

Program step	Input files	Output files
XYCORR	XDS.INP	XYCORR.LP X-CORRECTIONS.cbf Y-CORRECTIONS.cbf FRAME.cbf
INIT	XDS.INP X-CORRECTIONS.cbf Y-CORRECTIONS.cbf	INIT.LP BKGINIT.cbf BLANK.cbf GAIN.cbf
COLSPOT	XDS.INP X-CORRECTIONS.cbf Y-CORRECTIONS.cbf BLANK.cbf BKGINIT.cbf GAIN.cbf	COLSPOT.LP SPOT.XDS
IDXREF	XDS.INP SPOT.XDS	IDXREF.LP SPOT.XDS XPARM.XDS
DEFPIX	XDS.INP X-CORRECTIONS.cbf Y-CORRECTIONS.cbf BKGINIT.cbf XPARM.XDS	DEFPIX.LP BKGPIX.cbf ABS.cbf
XPLAN	XDS.INP XPARM.XDS BKGPIX.cbf X-CORRECTIONS.cbf Y-CORRECTIONS.cbf	XPLAN.LP
INTEGRATE	XDS.INP X-CORRECTIONS.cbf Y-CORRECTIONS.cbf BLANK.cbf BKGPIX.cbf GAIN.cbf XPARM.XDS	INTEGRATE.LP INTEGRATE.HKL FRAME.cbf
CORRECT	XDS.INP INTEGRATE.HKL REMOVE.HKL X-CORRECTIONS.cbf Y-CORRECTIONS.cbf	CORRECT.LP XDS_ASCII.HKL GXPARM.XDS DX-CORRECTIONS.cbf DY-CORRECTIONS.cbf *GX*-CORRECTIONS.cbf GY-CORRECTIONS.cbf ABSORP.cbf DECAY.cbf MODPIX.cbf

GAIN.cbf codes for the expected variation of the pixel contents in the background region of a data image. The variance of the contents of a pixel in the background region is GAIN·(pixel contents − detector noise). The variance is determined from the scatter of pixel values within a rectangular box (NBX=, NBY=) of size (2·NBX + 1)·(2·NBY + 1) centred at each image pixel in succession. The table GAIN.cbf is used to distinguish background pixels from 'strong' pixels that are part of a diffraction spot.

BKGINIT.cbf estimates the initial background at each pixel from a few data images specified by the input parameter BACKGROUND_RANGE=. The lookup table is obtained by adding the X-ray background from each image. Shaded regions on the detector (*i.e.* from the beamstop), pixels outside a user-defined circular region (TRUSTED_REGION=) or pixels with an undefined spatial correction value are classified as untrustworthy and marked by −3.

The following problem can be encountered in this step. Some detectors with insufficient protection from electromagnetic pulses may generate badly spoiled images whose inclusion leads to a completely wrong X-ray background table. These images can be identified in INIT.LP by their unexpected high mean pixel contents and this step should be repeated with a different set of images.

11.6.2.3. *COLSPOT*

COLSPOT locates strong diffraction spots occurring in a subset of the data images and saves their centroids in the file SPOT.XDS. The data subset is defined by contiguous image number ranges, where each range is specified by the keyword SPOT_RANGE=. As described in Chapter 11.3, spots are defined as sets of 'strong' pixels that are adjacent in three dimensions. The classification of 'strong' pixels is controlled by the decision constants STRONG_PIXEL= and BACKGROUND_PIXEL=. If the total number of 'strong' pixels occurring in the specified data images exceeds the upper limit as given by the input parameter MAXIMUM_NUMBER_OF_STRONG_PIXELS=, the weaker ones are discarded. A spot is accepted if it contains a minimum number of 'strong' pixels (MINIMUM_NUMBER_OF_PIXELS_IN_A_SPOT=) and if the spot centroid is sufficiently close to the location of the strongest pixel in the spot (SPOT_MAXIMUM-CENTROID=).

The following problem can be encountered in this step. Sharp edges such as ice rings in the images can lead to an excessive number of 'strong' pixels being erroneously classified as contributing to diffraction spots. These aliens could prevent *IDXREF* from recognizing the crystal lattice.

11.6.2.4. *IDXREF*

IDXREF uses the initial parameters describing the diffraction experiment as provided by XDS.INP and the observed centroids of the spots from the file SPOT.XDS to find the orientation, metric and symmetry of the crystal lattice, and refines all or a specified subset of these parameters [input parameter REFINE(IDXREF)=] . On return, the complete set of parameters are saved in the file XPARM.XDS and the original file SPOT.XDS is replaced by a file of identical name, now with indices attached to each observed spot. Spots not belonging to the crystal lattice are given indices 0, 0, 0. *XDS* considers the run to be successful if the coordinates of at least 70% of the given spots can be explained with reasonable accuracy (input parameter MAXIMUM_ERROR_OF_SPOT_POSITION=); otherwise, *XDS* will stop with an error message. Alien spots often arise because of the presence of ice or small satellite crystals and continuation of data processing may still be meaningful. In this case, *XDS* is called again with an explicit list of the subsequent steps specified in XDS.INP (input parameter JOB=DEFPIX XPLAN INTEGRATE CORRECT).

IDXREF uses the methods described in Kabsch (1993; Chapter 11.3) to determine a crystal lattice that explains the observed locations of the diffraction spots listed in the file SPOT.XDS. Firstly, a reciprocal-lattice vector referring to the unrotated crystal is computed from each observed spot centroid. Differences between any two reciprocal-lattice vectors that are above a specified minimal length (SEPMIN=) are accumulated in a three-dimensional histogram. These difference vectors will form clusters in the histogram, since there are many different pairs of reciprocal-lattice vectors of nearly identical vector difference. The clusters are found as maxima in the smoothed histogram (CLUSTER_RADIUS=) and a basis of three linearly independent cluster vectors is selected that allows all other cluster vectors to be expressed as nearly integral multiples of small magnitude with respect to this basis. The basis vectors and the 60 most populated clusters with attached indices are listed in `IDXREF.LP`. If many of the indices deviate significantly from integral values, the program is unable to find a reasonable lattice basis and all further processing will be meaningless.

If the space group and unit-cell parameters are specified, a reduced cell is derived and the reciprocal-basis vectors found above are reinterpreted accordingly; otherwise, a reduced cell is determined directly from the reciprocal basis. The parameters of the reduced cell, the coordinates of the reciprocal-basis vectors and their indices with respect to the reduced cell are reported.

Based on the orientation and metric of the reduced cell now available, *IDXREF* indexes up to 3000 of the strongest spots using the local-indexing method. This method considers each spot as a node of a tree and identifies the largest subtree of nodes which can be assigned reliable indices. The number of reflections in the ten largest subtrees is reported and usually shows a dominant first tree corresponding to a single lattice, whereas alien spots are found in small subtrees. Reflections in the largest subtree are used for initial refinement of the basis vectors of the reduced cell, the incident-beam wavevector and the origin of the detector, which is the point in the detector plane nearest to the crystal. Experience has shown that the detector origin and the direction of the incident beam are often specified with insufficient accuracy, which could easily lead to a misindexing of the reflections by a constant offset. For this reason, *IDXREF* considers alternative choices for the index origin and reports their likelihood of being correct. The parameters controlling the local indexing are INDEX_ERROR=, INDEX_MAGNITUDE=, INDEX_QUALITY= (corresponding to ε, φ and $1 - \ell_{\min}$ in Chapter 11.3) and INDEX_ORIGIN=h_0, k_0, l_0, which is added to the indices of all reflections in the tree. After initial refinement based on the reflections in the largest subtree, all spots which can now be indexed are included. Usually, the detector distance and the direction of the rotation axis are not refined, but if the spots were extracted from images covering a large range of total crystal rotation then better results are obtained by including these parameters in the refinement [REFINE(IDXREF)=] .

The refined metric parameters of the reduced cell are used to test each of the 44 possible lattice types as described in Chapter 11.3. For each lattice type, *IDXREF* reports the likelihood of its being correct and the conventional unit-cell parameters. The program step concludes with an overview of possible lattice symmetries, but makes no automatic decision for the space group. If the crystal symmetry is unknown, *XDS* will continue data processing with the crystal being described by its reduced-cell basis vectors and triclinic symmetry. Space-group assignment is postponed to the last program step, *CORRECT*, when integrated intensities are available.

The following problems can be encountered in this step.

(i) The indices of many difference-vector clusters deviate significantly from integral values. This can be caused by incorrect input parameters, such as rotation axis, oscillation angle or detector position, by a large fraction of alien spots in `SPOT.XDS`, by placing the detector too close to the crystal or by an inappropriate choice of the parameters SEPMIN= and CLUSTER_RADIUS= in densely populated images.

(ii) Indexing and refinement is unsatisfactory despite well indexed difference-vector clusters. This is probably caused by the selection of an incorrect index origin and *IDXREF* should be rerun with plausible alternatives for INDEX_ORIGIN= after a visual check of a data image with *XDS-Viewer*.

(iii) Despite successful indexing and refinement, *IDXREF* stops with the error message INSUFFICIENT PERCENTAGE OF INDEXED REFLECTIONS, complaining that less than 70% of the given spots could be explained. Alien spots often arise because of the presence of ice or small satellite crystals and continuation of data processing may still be meaningful. To continue data processing, just specify the missing processing steps in `XDS.INP` by JOB=DEFPIX XPLAN INTEGRATE CORRECT and call *XDS* again.

11.6.2.5. DEFPIX

DEFPIX recognizes regions in the initial background table (file `BKGINIT.cbf`) that are obscured by intruding hardware and marks the shaded pixels as untrusted. In addition, pixels that are outside a user-defined resolution range (INCLUDE_RESOLUTION_RANGE=) are marked and eliminated from the trusted region. The marked background table that is thus obtained is saved in the file `BKGPIX.cbf`, which is needed by the subsequent program steps. To recognize the obscured regions in the initial background, *DEFPIX* generates a control image (file `ABS.cbf`) that contains values of around 10 000 for unshaded pixels and lower values for shaded pixels. The classification of the pixels into reliable and untrusted pixels is based on the two input parameters VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= (default 6000 30 000) and INCLUDE_RESOLUTION_RANGE= (default 20.0 0.0). Pixels in the table `ABS.cbf` with a value outside the ranges specified by the two parameters are marked unreliable (by −3) in the background table `BKGPIX.cbf`.

The following problem can be encountered in this step. If the parameter VALUE_RANGE_FOR_TRUSTED_DETECTOR_PIXELS= specifies a value range that is too narrow, ‘good’ regions will erroneously be excluded from the trusted detector region. Check `BKGPIX.cbf` with the *XDS-Viewer* program and if necessary repeat the *DEFPIX* step with more appropriate values.

11.6.2.6. XPLAN

XPLAN supports the planning of data collection. It is based upon information provided by the input files `XPARM.XDS` and `BKGPIX.cbf`, both of which become available on processing a few test images with *XDS*. *XPLAN* estimates the completeness of new reflection data expected to be collected for each given starting angle and total crystal rotation and reports the results for a number of selected resolution shells in the file `XPLAN.LP`. To minimize the recollection of data, the name of a file containing already measured reflections can be provided by the input parameter REFERENCE_DATA_SET=.

The following problems can be encountered in this step.

(i) Incorrect results may occur for some space groups, *i.e.* $P4_2$, if the unit cell determined by *XDS* from processing a few test images implicates reflection indices that are inconsistent with those from the reference data set. However, the correct cell choice can be found by using the old data as a reference and repeating *CORRECT* with the appropriate reindexing transformation, followed by copying `GXPARM.XDS` to `XPARM.XDS`. The same applies if *IDXREF* was run for an unknown space group and then reindexed in *CORRECT*.

(ii) *XPLAN* ignores potential reflection overlap owing to the finite oscillation range covered by each image.

11.6.2.7. INTEGRATE

INTEGRATE determines the intensity of each reflection predicted to occur in the rotation data images (DATA_RANGE=) and saves the results in the file `INTEGRATE.HKL`.

The diffraction parameters needed to predict the reflection positions are initially provided by the file `XPARM.XDS`. These parameters are either kept constant or refined periodically using strong diffraction spots encountered in the data images. Whether refinement should be carried out at all and which parameters are to be refined can be specified by the user [input parameter REFINE(INTEGRATE)=]. The centroids of the strong spots in the data images are computed from pixels that exceed the background by a given multiple of standard deviations (input parameters SIGNAL_PIXEL=, BACKGROUND_PIXEL=). Strong spots are used in the refinement if their centroids are reasonably close to their calculated position (input parameter MAXIMUM_ERROR_OF_SPOT_POSITION=).

For determination of the intensity, approximate values describing the extension and the form of the diffraction spot must be specified. The shapes of all spots become very similar when the contents of each of their contributing image pixels are mapped onto a three-dimensional coordinate system, specific for each reflection, which has its origin on the surface of the Ewald sphere at the terminus of the diffracted beam wavevector (see Chapter 11.3). The transformed spot can roughly be described as a Gaussian involving two parameters: the standard deviations of the reflecting range σ_M (input parameter REFLECTING_RANGE_E.S.D.=σ_M) and the beam divergence σ_D (input parameter BEAM_DIVERGENCE_E.S.D.=σ_D). This leads to an integration region around the spot that is defined by the parameters δ_M (REFLECTING_RANGE=) and δ_D (BEAM_DIVERGENCE=), which are typically chosen to be 6–10 times larger than σ_M and σ_D, respectively. Appropriate values for these parameters are determined automatically by *XDS* (Chapter 11.3); the user has the option to override the automatic assignments.

Integration is carried out by a two-step procedure. In the first pass, spot templates are generated by superimposing the profiles of strong reflections after their mapping to the Ewald sphere. Grid points with a value above a minimum percentage of the maximum in the template (parameter CUT=) are marked for inclusion in the final integration. To allow for variations in their shape, profile templates are generated from reflections located at nine regions of equal size covering the detector surface and additional sets of nine to cover equally sized (parameter DELPHI=) batches of images. The actual integration is carried out in the second pass by profile fitting with respect to the spot shape determined in the first pass. Incomplete reflections below a minimum percentage of the observed reflection intensity (parameter MINPK=) will be discarded. Otherwise, the missing intensity is estimated from the learned reflection profiles.

On return from the *INTEGRATE* step, all spots expected to occur in the last data image are encircled and the modified image is saved as the file `FRAME.cbf` for inspection.

The following problems can be encountered in this step.

(i) Off-centred profiles indicate incorrectly predicted reflection positions by using the parameters provided by the file `XPARM.XDS` (*i.e.* misindexing by using a wrong origin of the indices), crystal slippage or change in the incident-beam direction.

(ii) Profiles extending to the borders of the box indicate too-small values of the parameters BEAM_DIVERGENCE= or REFLECTING_RANGE=. This leads to incorrect integrated intensities because of truncated reflection profiles and unreliable background determination.

(iii) Display of the file `FRAME.cbf` shows spots which are not encircled. If these unexpected reflections are not close to the spindle and are not ice reflections, then it is likely that the parameters provided by the file `XPARM.XDS` are wrong.

11.6.2.8. CORRECT

CORRECT applies correction factors to the intensities and standard deviations of all reflections found in the file `INTEGRATE.HKL`, determines the space group if unknown and refines the unit-cell parameters, reports the quality and completeness of the data set and saves the final integrated intensities in the file `XDS_ASCII.HKL`. Some of the employed algorithms are new and are described in Chapter 11.3.

CORRECT accepts reflections from the file `INTEGRATE.HKL` that are

(i) recorded (parameter MINPK=) on specified images (parameter DATA_RANGE=);

(ii) within a given resolution range (parameter INCLUDE_RESOLUTION_RANGE=);

(iii) outside ice rings (parameter EXCLUDE_RESOLUTION_RANGE=);

(iv) not overloaded (parameter OVERLOAD=); and

(v) not marked for exclusion in the file `REMOVE.HKL`.

Thus, the user has the option to exclude unreliable reflections from the final data set by repeating the *CORRECT* step with appropriate parameter values.

The intensities of the accepted reflections are first corrected for effects arising from polarization of the incident beam (parameters FRACTION_OF_POLARIZATION=, POLARIZATION_PLANE_NORMAL=) and absorption effects (parameters AIR=, SILICON=, SENSOR_THICKNESS=) arising from differences in path lengths of the diffracted beam. These corrections do not depend on knowledge of the space group.

The integrated intensities of the reflections in the file `INTEGRATE.HKL` may or may not have been indexed in the correct space group; for the purpose of integration, it is important only that all reflections occurring in the data images have been indexed with respect to some unit-cell basis and that their locations on the images were hit exactly. The correct reflection indices in the true space group are always a linear transformation of the original indices used in `INTEGRATE.HKL`. All lattices consistent with the locations of the reflections saved in `INTEGRATE.HKL` (decision parameters MAX_CELL_AXIS_ERROR=, MAX_CELL_ANGLE_ERROR=) and their corresponding linear transformations are printed to provide a useful overview similar to that shown in `IDXREF.LP`.

If the space group is not specified, *XDS* proposes one of the enantiomorphous space groups without screw axes that is compatible with the observed lattice symmetry and explains the intensities of a subset of the reflections (parameter TEST_RESOLUTION_RANGE=) at an acceptable $R_{\rm meas}$ (Diederichs & Karplus, 1997; Weiss, 2001) using a minimum number of unique reflections. The criteria for an acceptable $R_{\rm meas}$ are controlled by the decision parameters MIN_RFL_Rmeas= and MAX_FAC_Rmeas=.

The user can always override the automatic decisions by specifying the correct space-group number (parameter SPACE_GROUP_NUMBER=) and unit-cell parameters (parameter UNIT_CELL_CONSTANTS=) in `XDS.INP` and repeating the *CORRECT* step. This provides a simple way to rename orthorhombic unit-cell parameters, which often becomes necessary if screw axes are present. In addition, the user has the option to specify the following in `XDS.INP`:

(i) a reference data set (parameter REFERENCE_DATA_SET=),
(ii) a reindexing transformation (parameter REIDX=) and
(iii) three basis vectors if known from processing a previous data set taken at the same crystal orientation in a multi-wavelength experiment (parameters UNIT_CELL_A-AXIS=, UNIT_CELL_B-AXIS=, UNIT_CELL_C-AXIS=).

The possibility of comparing the new data with a reference data set is particularly useful for resolving the issue of alternative settings of polar or rhombohedral cells (such as $P4$, $P6$ and $R3$). Also, reference data are quite useful for recognizing misindexing or for testing potential heavy-atom derivatives.

For refinement of the unit-cell parameters [parameter REFINE(CORRECT)=], *CORRECT* uses a subset of the accepted reflections whose observed centroid is sufficiently close to the predicted spot position (parameter MAXIMUM_ERROR_OF_SPOT_POSITION=). The refined set of parameters is saved in the file `GXPARM.XDS`, which has an identical layout to the file `XPARM.XDS` produced by *IDXREF*. If the crystal has not slipped during data collection, these parameters are quite accurate.

Other correction factors (parameter CORRECTIONS=) which partially compensate for radiation damage, absorption effects and variations in the sensitivity of the detector surface are determined from the symmetry-equivalent reflections usually found in the data images. The corrections are chosen such that the integrated intensities of symmetry-equivalent reflections come out as similar as possible. The user may control application of the various corrections by specifying the parameter CORRECTIONS= by a combination of the keywords DECAY MODULATION ABSORPTION. Whether Friedel pairs are considered as symmetry-equivalent reflections in the calculation of the correction factors depends on the values of the two parameters STRICT_ABSORPTION_CORRECTION= and FRIEDEL'S_LAW=. The number of correction factors is controlled by the input parameters MINIMUM_I/SIGMA=, NBATCH= and REFLECTIONS/CORRECTION_FACTOR=.

The residual scatter in intensity of symmetry-equivalent reflections is used to estimate their standard deviations. Here, the initial estimate $v_0(I)$ (obtained from the *INTEGRATE* step) for the variance of the reflection intensity I is replaced by $v(I) = a[v_0(I) + bI^2]$. The two constants a and b are chosen to minimize discrepancies between $v(I)$ and the variance estimated from sample statistics of symmetry-related reflections. Based on the more realistic error estimates for the intensities, outliers are recognized by comparison with other symmetry-equivalent reflections. These outliers are included in the main output file `XDS_ASCII.HKL`, in which they are marked by a negative sign attached to the estimated standard deviations of their intensity. Classification of a reflection as a misfit is controlled by a decision constant which has the default value of WFAC1=1.5. Specification of a lower value such as WFAC1=1.0 by the user will lead to an increasing number of misfits and lower R factors as outliers are not included in the reported statistics.

Data quality as a function of resolution is described by the agreement of intensities of symmetry-related reflections and quantified by the R factors $R_{\rm merge}$ and the more robust indicator $R_{\rm meas}$ (Diederichs & Karplus, 1997; Weiss, 2001). These R factors as well as the intensities of all reflections with indices of type $h00$, $0k0$ and $00l$ and those expected to be systematically absent provide important information for identification of the correct space group. Clearly, large R factors or many rejected reflections or large observed intensities for reflections that are expected to be systematically absent suggest that the assumed space group or indexing is incorrect. The presence or absence of anomalous scatterers is specified by the parameter FRIEDEL'S_LAW=. Finally, *CORRECT* analyzes the distribution of reflection intensities as a function of their resolution and reports outliers from the Wilson plot. Often, these aliens arise from ice rings in the data images. To suppress the unwanted reflections from the final output file `XDS_ASCII.HKL`, the user copies them to a file named `REMOVE.HKL` in the current directory and repeats the *CORRECT* step.

The following problems can be encountered in this step.

(i) Incomplete data sets may lead to wrong conclusions about the space group, as some of its symmetry operators might not be involved in the R-factor calculations.
(ii) Often, the *CORRECT* step is repeated several times. It should be remembered that *XDS* overwrites earlier versions of the output files `XDS_ASCII.HKL`, `GXPARM.XDS` *etc.*

11.6.3. *XSCALE*

The scaling program *XSCALE*

(i) puts one or more files obtained from data processing with *XDS* on a common scale and reports the completeness and quality of the data sets;
(ii) offers a choice of either combining symmetry-equivalent observations into a single unique reflection or saving the scaled but unmerged observations in the output file;
(iii) allows several output files that are placed on the same scale, a feature that is recommended for multiwavelength anomalous diffraction (MAD) data sets taken from the same crystal at different wavelengths;
(iv) determines correction factors that partially compensate for absorption effects, sensitivity variations in the detector plane and radiation damage; and
(v) can correct reflections individually for radiation damage (Diederichs *et al.*, 2003).

The program uses a new fast algorithm (Chapter 11.3) and imposes no limitations on the number of data sets or scaling/correction factors. The easiest way to run *XSCALE* is to copy a template input file named `XSCALE.INP` to a new directory and to replace the parameter values by the appropriate values describing the actual scaling run. The input parameters may be given in arbitrary order, except for the parameters defining the input and output reflection files (INPUT_FILE=, OUTPUT_FILE=). Here, an output file is defined first by the parameter OUTPUT_FILE= that will include the scaled and merged reflections from all following input files specified by the parameters INPUT_FILE= until the next occurrence of OUTPUT_FILE= in `XSCALE.INP`. An arbitrary number of output files can be specified (together with their set of input files) in a single run of *XSCALE*. All output files are then on the same scale, which is a useful program feature for MAD data sets.

The reflections in each output file will be unmerged and Friedel pairs will be considered to be different if this holds for all of the input data sets unless explicitly redefined by the parameters MERGE= and FRIEDEL'S_LAW=. Moreover, each output file accepts an additional parameter that controls how the Friedel pairs of the input files are treated in the calculation of the absorption correction factors. If STRICT_ABSORPTION_CORRECTION=FALSE, Friedel pairs are treated as symmetry-equivalent reflections in these calculations, which could lead to an underestimate of the anomalous differences in the presence of anomalous scatterers. Friedel pairs are only treated as different reflections in the calculations if STRICT_ABSORPTION_CORRECTION=TRUE and FRIEDEL'S_LAW=FALSE.

For each input file, a resolution window for accepting reflections (INCLUDE_RESOLUTION_RANGE=), the extent of absorption corrections (CORRECTIONS=DECAY MODULATION ABSORPTION) and the number of correction factors (NBATCH=) can be specified. Finally, each input data set can be corrected for radiation damage by specifying the name of the crystal the data set was obtained from (CRYSTAL_NAME=). Specification of this parameter implicates zero-dose extrapolation of individual reflection intensities to compensate for the effects of radiation damage experienced by the crystal so far (see Diederichs *et al.*, 2003).

Each resulting scaled data set is of XDS_ASCII format. It can be converted into a CCP4-style multi-record MTZ file using the copy feature of the program *POINTLESS* (Evans, 2006) available from the web (ftp://ftp.ccp4.ac.uk/ccp4/6.0.2/prerelease/pointless.html) or converted by *XDSCONV* into the format required by various structure-solution packages.

11.6.4. *XDSCONV*

XDSCONV accepts reflection-intensity data files as produced by *XSCALE* or *CORRECT* and converts them into the format required by software packages for structure determination. *XDSCONV* estimates structure-factor moduli based on the assumption that the intensity data set obeys Wilson's distribution and uses a Bayesian approach to statistical inference as described by French & Wilson (1978). The output file generated may inherit the test reflections previously used to calculate a free *R* factor (Brünger, 1992) or may contain new test reflections selected by *XDSCONV*.

11.6.5. Parallelization of *XDS*

In order to efficiently use modern multiprocessor hardware, a major effort has been undertaken to replace the original code of *XDS* by routines that can run concurrently with very little need for synchronization. As described above, data processing by *XDS* is organized into eight steps that must be executed in a fixed order since the result of each step is needed as input for the subsequent ones. Thus, the only way to speed up processing is to make each step faster.

The most computationally intensive steps are *COLSPOT* and *INTEGRATE* and, to a lesser degree, the routine that refines diffraction parameters in *IDXREF* and *CORRECT*. Thus, the highest savings in wall clock time are expected to result from changing these routines so that each one can make efficient use of the multiprocessor hardware. Two methods can be used (simultaneously) to speed up data processing.

In the first method, *XDS* divides the set of data images into approximately equal portions, calls a shell script that starts an independent job for processing each portion of images by the computer cluster and waits until all jobs have finished. The number of such independent jobs can be limited by the user (MAXIMUM_NUMBER_OF_JOBS=); up to 99 jobs are allowed. This method works even if the processors do not share the same address space, since the jobs are independent processes that do not communicate at all.

The second method uses OpenMP to control execution by a team of threads and relies on a shared-memory multiprocessor platform. This allows the program to exploit data parallelism at a more fine-grained level to speed up refinements and routines for setting up and solving systems of linear equations. The maximum number of threads that can be employed by the parallel version of *XDS* (xds_par, xscale_par) can be limited by the user (MAXIMUM_NUMBER_OF_PROCESSORS=); up to 32 processors can be used. OpenMP has been chosen for execution control because it hardly adds to the complexity of the program code and most importantly does not require the maintenance of separate versions of the source code depending on whether the program is intended for execution by a team of processors or just by a single CPU. Moreover, OpenMP has become the *de facto* standard and compilers accepting OpenMP directives are available for most shared-memory multiprocessor platforms.

The new version of *COLSPOT* comprises an initial part, a concurrent procedure and a final part. After initialization each available processor is kept busy analysing its share of rotation images for strong pixels, which are saved in a processor-specific file. In the final sequential part of *COLSPOT* all files resulting from the concurrent computations are read and the location addresses, image running numbers and signal values of the strong pixels are stored in a hash table. Strong pixels belonging to the same spot can be located rapidly in this table and the centroids of the spots are saved in the final output file from this step.

For the *INTEGRATE* step, the rotation images are divided into approximately equal portions for independent processing under control by a shell script according to the first method described above. When all jobs have finished, the integrated intensities from each independent file are joined. Minor problems could occur for reflections that receive intensity contributions from images that have been processed by different jobs. Compared with processing as a single job, the observed intensity differences are small and disappear if the different jobs use identical reference profiles and diffraction parameters to predict spot locations [to avoid refinements, specify REFINE(INTEGRATE)=!].

In addition, each of the independent jobs can be executed by a team of processors controlled by OpenMP. The rotation images analysed by each job are split into a sequence of batches of consecutive images that cover a total rotation range that is large enough to accommodate the integration domain. The batches are evaluated in strictly sequental order; parallel processing is confined to images within each batch.

The restructured routine for the *INTEGRATE* step consists of code regions for parallel execution interspersed by sequential sections. After initialization, strong reflections and their mean size and extent are determined concurrently. The diffraction parameters are refined in parallel-processing mode based on the observed spot locations. In the following sequential section a database is generated containing information about all reflections occurring in this batch of images. A subset of strong reflections is also identified that is useful for the subsequent reflection-profile learning pass. The mean profile of these reflections is determined concurrently in a second pass through the images in the batch. Reflection integration by profile fitting is carried out in parallel in the third cycle through the batch. In the final sequential step the results from each job, which have been saved in files, are harvested and intensity contributions to the same reflection from adjacent batches are merged.

11.6.6. Availability

Documentation and executable versions of the *XDS* package for widely used computer systems running under Linux or OSX can be obtained from the *XDS* homepage (http://xds.mpimf-heidelberg.mpg.de/) free of charge for use by academics for noncommercial applications. Additional information can be found at http://strucbio.biologie.uni-konstanz.de/xdswiki/index.php/XDS.

For looking at rotation data images and control images generated by *XDS*, an open-source program *XDS-Viewer* written by Michael Hoffer can be obtained from http://xds-viewer.sourceforge.net under the GNU General Public License.

A graphical interface *XDSi* (Kursula, 2004) is available (http://cc.oulu.fi/~pkursula/xdsi.html) that simplifies the operation of *XDS*.

This chapter is based on a recent article published in *Acta Crystallographica Section D* (Kabsch, 2010).

References

Bernstein, H. J. & Ellis, P. J. (2005). *CBFlib: an ANSI C library for manipulating image data*. In *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 544–556. Heidelberg: Springer.

Bernstein, H. J. & Hammersley, A. P. (2005). *Specification of the Crystallographic Binary File (CBF/imgCIF)*. In *International Tables for Crystallography*, Vol. G, edited by S. R. Hall & B. McMahon, pp. 37–43. Heidelberg: Springer.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Diederichs, K. & Karplus, P. A. (1997). *Improved R-factors for diffraction data analysis in macromolecular crystallography*. *Nat. Struct. Biol.* **4**, 269–274.

Diederichs, K., McSweeney, S. & Ravelli, R. B. G. (2003). *Zero-dose extrapolation as part of macromolecular synchrotron data reduction*. *Acta Cryst.* D**59**, 903–909.

Evans, P. (2006). *Scaling and assessment of data quality*. *Acta Cryst.* D**62**, 72–82.

French, S. & Wilson, K. (1978). *On the treatment of negative intensity observations*. *Acta Cryst.* A**34**, 517–525.

Kabsch, W. (1988*a*). *Automatic indexing of rotation diffraction patterns*. *J. Appl. Cryst.* **21**, 67–71.

Kabsch, W. (1988*b*). *Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector*. *J. Appl. Cryst.* **21**, 916–924.

Kabsch, W. (1993). *Automatic processing of rotation diffraction data from crystals of initially unknown symmetry and cell constants*. *J. Appl. Cryst.* **26**, 795–800.

Kabsch, W. (2010). *XDS*. *Acta Cryst.* D**66**, 125–132.

Kursula, P. (2004). *XDSi: a graphical interface for the data processing program XDS*. *J. Appl. Cryst.* **37**, 347–348.

Weiss, M. S. (2001). *Global indicators of X-ray data quality*. *J. Appl. Cryst.* **34**, 130–135.

International Tables for Crystallography (2012). Vol. F, Chapter 11.7, pp. 311–316.

Chapter 11.7. Detecting twinning by merohedry

T. O. Yeates and Y. Tsai

11.7.1. Introduction

Twinning is a growth disorder in which different regions or domains of a crystal specimen are oriented differently with respect to each other (reviewed in Yeates, 1997; Dauter, 2003; Parsons, 2003; Helliwell, 2008). Numerous types of twinning are possible and a hierarchy among the types has been articulated (Yeates & Fam, 1999) (Fig. 11.7.1.1). In the most general type, differently oriented twin domains in the crystal have lattices that are distinguishable from each other. As a result, the different domains are associated with distinct reciprocal lattices and therefore with distinguishable interpenetrating diffraction patterns. This type of twinning is therefore generally recognized in X-ray diffraction images. The situation is referred to as twinning by non-merohedry and has also been referred to occasionally in the literature as 'epitaxial twinning'. However, the latter term is usually used to describe the growth of a crystal of heterogeneous molecules. Treatment of twinning by non-merohedry usually amounts to a data-collection problem, as only a fraction of the diffraction spots in a given image correspond to a single lattice. A further complication can arise from the presence of a subset of reflections (*e.g.* those lying on an axis of rotation that relates the twin domains) whose multiple distinct contributions from the different domains overlap in the diffraction pattern. Strategies for unit-cell determination and reflection integration have been developed and implemented in various X-ray data processing programs (Sparks, 1999; Sheldrick, 1997, 2002, 2004, 2008; Duisenberg, 1992; Herbst-Irmer & Sheldrick, 1998; Duisenberg *et al.*, 2003; Dix *et al.*, 2007). Those programs have been developed mainly in the context of small-molecule applications, but are often suitable for problems involving macromolecular crystals as well. We do not consider twinning by non-merohedry further in this chapter, but focus instead on a special category (Fig. 11.7.1.1) known as twinning by merohedry.

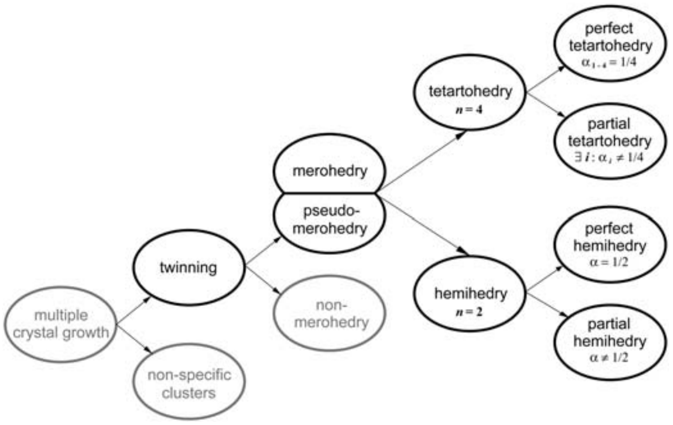


Figure 11.7.1.1
Hierarchy of multiple types of twinning. Merohedry is elevated over non-merohedry as it results in the exact superposition of diffraction patterns. Tetartohedry is ranked higher than hemihedry because it involves a larger number of domain orientations. Perfect merohedry is given higher rank than partial merohedry as it results in higher apparent symmetry than exists in the crystal. Darker print highlights the scope of the text. (Modified from Yeates & Fam, 1999.)

The synonymous term 'merohedral twinning' is also widely used in the macromolecular literature, although it has been noted that the latter usage introduces potential confusion in that the adjective 'merohedral', or '*mérohèdre*' (Friedel, 1926), is used in mathematical crystallography in a somewhat different context as a specific property of certain symmetries.

11.7.2. Twinning by merohedry – considerations of lattice symmetry

In twinning by merohedry, the lattices of the differently oriented domains are superimposable, although the operations that relate the domains are not operations of the space-group symmetry (Fig. 11.7.2.1). This is possible only in special symmetries, namely those where the rotational symmetry of the lattice exceeds the underlying rotational symmetry of the crystal space group; we consider only rotational operations here, in order to focus on situations relevant to biological macromolecules. Symmetries where this occurs are listed in Table 11.7.2.1; they include crystal point groups 4, 3, 312, 321, 6 and 23. Space group *P*4 provides a simple example. It requires a tetragonal lattice, and that lattice contains twofold axes of rotational symmetry (perpendicular to the fourfold axis) which are not utilized by the *P*4 symmetry of the crystal. Therefore, two separate crystal domains, each obeying *P*4 symmetry, can grow together within the same specimen in a way such that they are rotated with respect to each other by one of these twofold symmetry axes of the tetragonal lattice. The operation used to describe the rotation is referred to as the twin operation. Because the lattices of the two domains are indistinguishable, the separate domains give rise to diffraction patterns that exactly superimpose in reciprocal space. Therefore, multiple distinct reflections (*e.g.* one from each of the two twin domains) fall precisely on top of each other (Fig. 11.7.2.1). Because warning signs are usually not present during inspection of diffraction patterns, twinning by merohedry must be detected

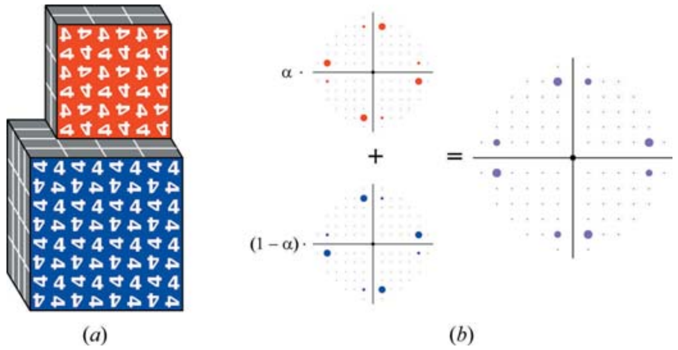


Figure 11.7.2.1
A cartoon depicting partial merohedral twinning in space group *P*4. (*a*) Two twin domains growing together, related by a twofold twin operation perpendicular to the fourfold symmetry axis. (*b*) The diffraction patterns of individual domains and their overlapping combination expected in a diffraction experiment. (Adapted from Yeates & Fam, 1999.)

Copyright © 2012 International Union of Crystallography

Table 11.7.2.1
Symmetries in which twinning by merohedry can occur in biological macromolecules

Point group	Space groups	Apparent point group when $\alpha = 1/2$
4	$P4$ $P4_1$ $P4_2$ $P4_3$ $I4$ $I4_1$	422
3	$P3$ $P3_1$ $P3_2$ $R3$†	6 or 321 or 312‡
312	$P312$ $P3_112$ $P3_212$	622
321	$P321$ $P3_121$ $P3_221$	622
6	$P6$ $P6_1$ $P6_2$ $P6_3$ $P6_4$ $P6_5$	622
23	$P23$ $P2_13$ $I23$ $I2_13$ $F23$	432

† Excluding cases of reticular twinning, $R3$ can only be twinned towards $R32$. ‡ Point group 3 gives apparent symmetry 622 under perfect tetartohedral symmetry ($\alpha_{1\text{–}4} = 1/4$).

after data collection, and special steps must be taken during structure determination if unaffected crystals cannot be grown.

The case of space group $P4$ described above is an example of the simplest type of twinning by merohedry, namely where only two distinct domain orientations are possible. The case of two domain orientations ($n = 2$) is referred to as twinning by hemihedry. We focus mainly on hemihedry in this chapter but provide brief comments on higher-order cases, such as tetartohedry ($n = 4$), at the end. Another distinct category of twinning arises in crystals where the lattice very nearly obeys some higher rotational symmetry, not by virtue of a strict requirement, as in $P4$ and its tetragonal lattice, but because of a fortuitous geometry of the unit cell. These cases are referred to as twinning by pseudo-merohedry. A simple example is provided by a monoclinic crystal with the unique β angle very nearly equal to 90°. The lattice is nearly orthorhombic, in which it is possible to have two domain orientations related by a twofold rotation about one of the principal axes perpendicular to the crystallographic axis of twofold symmetry. Situations of twinning by pseudo-merohedry are relatively common (Rudolph *et al.*, 2004; Hamdane *et al.*, 2009). The method of treatment is similar to those of typical merohedry, but detection requires special attention since the phenomenon depends on specific unit-cell geometries. Modern software programs are able to detect cases where twinning by pseudo-merohedry is possible by identifying potentially higher metric symmetry within an apparently lower-symmetry unit cell (Adams *et al.*, 2002, 2004; Zwart *et al.*, 2005*b*). Yet another special case is reticular twinning, wherein only a subset of reflections are superimposed on others by the twinning operation while the remaining reflections are not overlapped (Herbst-Irmer & Sheldrick, 2002). The simplest case is space group $R3$, twinned about a twofold axis along **c** (for example); reflections for which $l = 3n$ overlap with other reflections in the same layer under the twin operation, while reflections in the other layers do not overlap. To date, reticular twinning is not well documented in macromolecular crystals and is not covered further here.

11.7.3. Considerations of length scale and effects in reciprocal space

Twinning is distinguished from other types of disorder by the length scale over which the disorder occurs. In cases of twinning, individual twin domains behave as ideal crystals whose sizes are presumed to be large compared with the coherence length of the X-ray beam. Therefore, the scattering from twinned specimens behaves as a sum of intensities from the separate twin domains, rather than as a sum of complex-valued structure factors; the X-ray scattering from one domain is presumed not to interfere with the scattering from the other domain(s). The organization of differently oriented domains in twinned crystal specimens has been studied in select cases (Efremov *et al.*, 2004; Ko *et al.*, 2001), but is generally not well understood for macromolecular crystals. Based on the behaviour of twinned crystals of small molecules, various types of physical arrangements are likely.

11.7.4. Extent of twinning: the twin fraction

The twin fraction describes the fractional volumes of the different domain orientations within a crystal specimen. In the case of twinning by hemihedry, just one twin fraction, α, needs to be specified. Its value is taken to be the minor component, meaning a value less than or equal to 1/2. The twin fraction dictates the relative weights with which two twin-related reflections contribute to each observed intensity. The basic system of equations is

$$\begin{aligned} I_{\rm obs}(\mathbf{h}_1) &= (1-\alpha)I(\mathbf{h}_1) + \alpha I(\mathbf{h}_2), \\ I_{\rm obs}(\mathbf{h}_2) &= \alpha I(\mathbf{h}_1) + (1-\alpha)I(\mathbf{h}_2), \end{aligned} \qquad (11.7.4.1)$$

where $\mathbf{h}_1$ and $\mathbf{h}_2$ are twin-related reflections, I represents the true crystallographic intensity of a reflection (*i.e.* the square of the structure-factor magnitude) and $I_{\rm obs}$ represents an observed intensity (Fig. 11.7.2.1).

The value of the twin fraction has a critical effect on how twinning manifests itself in practice. Two essentially distinct scenarios arise. First, if the twin fraction is significantly less than 1/2, a situation typically referred to as 'partial twinning' occurs, and partial higher symmetry may be evident during data processing and analysis. For example, if the true space group is $P4$ and the crystal specimen is partially twinned, one might find that the merging R value is lower than random when attempts are made to reduce the data under the higher tetragonal symmetry $P422$. Similarly, when the data set reduced in $P4$ is analysed using a self-rotation function, partial higher symmetry may be evident along what would be symmetry axes in $P422$. However, such symptoms can also be consistent with an alternative explanation, namely local or noncrystallographic symmetry (NCS) nearly coincident with potentially higher crystallographic symmetry ($P422$ in the present example). Discriminating partial twinning from pseudo-crystallographic NCS is therefore a key problem in analysing data in space groups that support twinning.

In the second scenario, the twin fraction α is very nearly 1/2, giving rise to a situation typically referred to as 'perfect twinning'. In that case, pairs of twin-related observations are necessarily equal. The recorded data set therefore obeys higher symmetry. For example, if a crystal in $P4$ is perfectly twinned by merohedry, the data set will obey 422 symmetry. The challenge in a case of perfect twinning is to recognize that the correct space-group symmetry is lower than that implied by the observed intensity data.

11.7.5. Indications of twinning

Unusual crystal morphology sometimes provides an early suggestion of twinning. In particular, concave crystals or crystals having 're-entrant' faces are suspect. However, most cases of twinning in macromolecules are not identified until structure determination is in progress. It is relatively rare, but in some cases twinning can be identified decisively when an impossibly dense molecular packing would be required to generate a crystal with symmetry as high as that implied by the observed data (Redinbo & Yeates, 1993). Sometimes, the possibility of twinning is raised during atomic refinement when, for unexplained reasons, the refinement R values cannot be brought down to acceptable limits. However, it should be emphasized that problematic refinement is never sufficient evidence by itself to conclude that a crystal is twinned; instead, a careful analysis of the observed data is required. Modern data-reduction programs now incorporate various statistical analyses for this purpose, making it relatively straightforward to identify the presence of twinning before proceeding further with structure determination.

Twinning has two related but distinct effects on intensity data. First, the observed intensities of twin-related reflections tend to be more similar to each other than expected for independent reflections; in the limit of perfect twinning they become equal. Second, the overall statistical distribution of intensities is affected, so that classic Wilson statistics are not obeyed. Therefore, two essentially different kinds of tests exist to analyse intensity statistics for the presence of twinning.

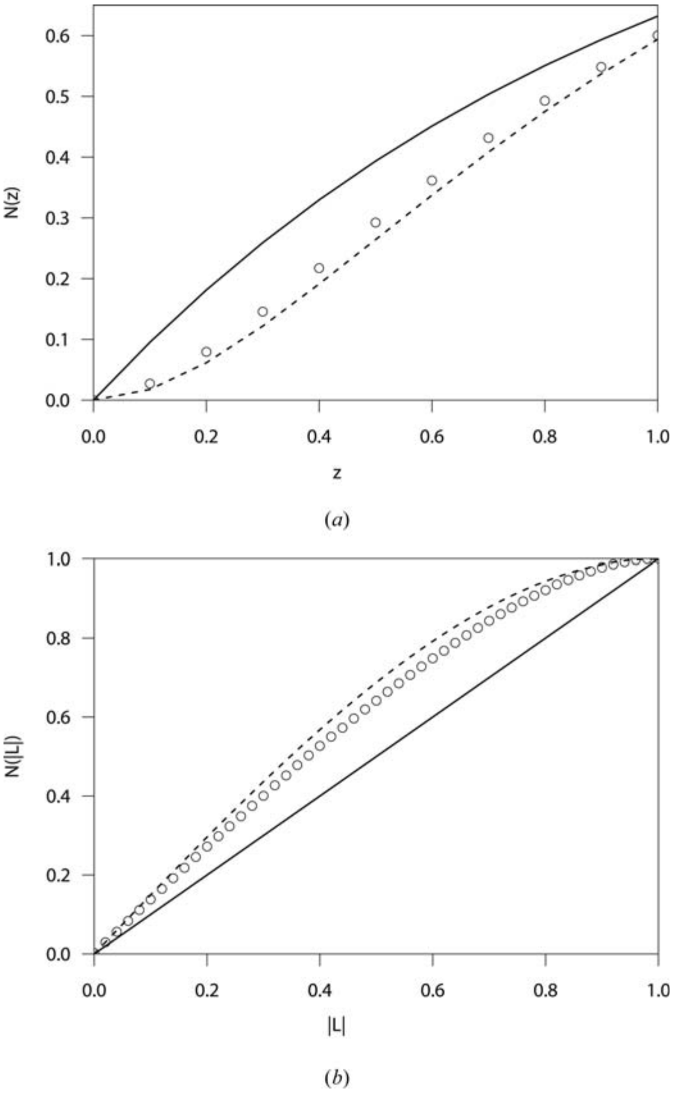


Figure 11.7.6.1
Detection of high or perfect twinning by analysing overall intensity statistics. (*a*) A traditional examination of cumulative intensity statistics. z is the normalized intensity and $N(z)$ is its cumulative distribution. Theoretical curves are shown for perfectly twinned (dashed line) and untwinned acentric data (solid line). The experimental curve (open circles) is for the protein Dicer [Protein Data Bank (PDB; Berman *et al.*, 2000) code 2qvw; MacRae & Doudna, 2007], which is essentially perfectly twinned. (*b*) A plot of the local intensity difference, L (see text). Theoretical curves for perfectly twinned (dashed line) and untwinned (solid line) data are shown. Data from a perfectly twinned crystal, VP5CT (Padilla & Yeates, 2003), are plotted with open circles. The crystal suffered from pseudo-centring, which did not interfere with the interpretation of twinning using local intensity statistics; reflection pairs were chosen to have indices differing by even numbers.

11.7.6. Twinning tests based on overall intensity statistics

Twinning causes a shift in the typical intensity distribution from an exponential to a sigmoidal shape (Fig. 11.7.6.1*a*). Various measures of the intensity distribution have been used to diagnose twinning. The Wilson ratio, $\langle F\rangle^2/\langle F^2\rangle$ (where the angle brackets denote the average), is 0.785 for ordinary acentric data and 0.885 for perfectly twinned data (Stanley, 1972; Wilson, 1949). Similarly, the value of $\langle|E^2-1|\rangle$ (where E^2 is defined as $F^2/\langle F^2\rangle$) is 0.736 for ordinary acentric data and 0.541 for perfectly twinned data (Bruker, 2000). The moments of the intensity distribution have simple, rational values. For example, the ratio $\langle I^2\rangle/\langle I\rangle^2$, sometimes referred to as the Stanley factor, is 2.0 for ordinary acentric data and 1.5 for perfectly twinned data (Stanley, 1972). For partial twinning, intermediate behaviour is observed and complex expressions for the resulting intensity distributions have been worked out (Rees, 1980). A key difficulty with these statistical tests is that they assume that the members of the set of reflections being evaluated all have the same expected value. If the diffraction is nearly isotropic, this condition can be satisfied relatively easily by normalizing the reflection intensities in thin resolution shells. In contrast, anisotropy presents a problem that must be corrected if these traditional tests are to be useful. If the anisotropy has a relatively simple form (*e.g.* if it follows an elliptical Gaussian function), then correction may be straightforward, but otherwise correction can remain problematic. Crystals in which distinct molecules in the asymmetric unit are related by translational shifts create similar problems: classes of strong and weak reflections are present in a single resolution shell. Cases of anisotropy and translational pseudo-symmetry both give intensity distributions that are shifted counter to the shift caused by twinning, and can therefore mask its presence.

One method for examining overall intensity statistics in a way that seeks to overcome the challenges described above has been described (Padilla & Yeates, 2003). Reflections near each other in reciprocal space are considered in pairs. The difference between the two intensities divided by their sum is taken as the quantity to be analysed,

$$L \equiv \frac{(I_{\mathrm{obs},A} - I_{\mathrm{obs},B})}{(I_{\mathrm{obs},A} + I_{\mathrm{obs},B})}, \qquad (11.7.6.1)$$

where L is the so-called local intensity difference, and $I_{\mathrm{obs},A}$ and $I_{\mathrm{obs},B}$ are the observed intensities of two reflections that are not related by any crystal symmetry or potential twin operation. Choosing the reflection pairs to be near in reciprocal space ensures that anisotropy effects will not affect their expected values differently. Countering translational pseudo-symmetry is a more difficult problem, but common cases of pseudo-centring or unit-cell doubling can be treated effectively by requiring that reflection pairs have the same parity (*i.e.* differ by an even value for all three indices). In principle, more complex types of pseudo-symmetry can be handled with additional care (Zwart *et al.*, 2005*a*).

For two ordinary (untwinned) acentric intensities following Wilson statistics, $|L|$ has a uniform probability distribution from 0 to 1. For perfectly twinned intensities, the distribution of L has a quadratic form. An observed data set can therefore be analysed and the cumulative distribution of $|L|$ compared with simple theoretical curves (Fig. 11.7.6.1*b*). Likewise, the expected values and moments of L are simple fractions for both untwinned and

Table 11.7.6.1
Twinning statistics for acentric intensity data from a crystal specimen with different numbers of twin domain orientations (n)

Values are given for crystals exhibiting no twinning (n = 1), perfect twinning by hemihedry (n = 2) and perfect twinning by tetartohedry (n = 4).

n	$\langle\lvert L\rvert\rangle$	$\langle L^2\rangle$	$\langle I^2\rangle/\langle I\rangle^2$
1	1/2	1/3	2
2	3/8	1/5	3/2
4	35/128	1/9	5/4

perfectly twinned cases (Table 11.7.6.1). Values for $\langle|L|\rangle$ are 1/2 and 3/8 for untwinned and perfectly twinned data, respectively.

When applying tests of overall intensity statistics, situations of partial twinning produce values intermediate between those expected for untwinned and perfectly twinned data. The twinning fraction can therefore be estimated by such tests. However, owing to systematic deviations of the types noted above, this is generally not the best way to estimate α. In particular, if α is small the presence of partial twinning can be missed. The strength of these tests is in identifying cases of high or perfect twinning. Consequently, an examination of overall intensity statistics is sometimes referred to as a test for perfect twinning. Such tests typically provide decisive support for a claim of high or perfect twinning.

11.7.7. Tests for partial twinning based on comparison of twin-related reflections

Once model-derived intensities become available during structure determination, the presence of partial twinning can be detected reliably and α can be estimated with high precision (see Chapter 18.12). However, at the outset of structure determination, identifying partial twinning unambiguously and estimating α are more challenging problems, and numerous methods have been developed. One method, referred to as a Britton plot and popularized by Fisher & Sweet (1980), is based on the behaviour of equation (11.7.4.1). Given a pair of twin-related observed intensities and a presumed value for α, the two underlying crystallographic intensities can be calculated by inverting the linear equations. If α is chosen to be higher than its correct value, this process leads to some of the crystallographic intensities having negative calculated values. The correct value of α can therefore be estimated by plotting the number of negative calculated intensities obtained as a function of α. For error-free data, the correct value for α is the highest value that gives no negative calculated intensities.

Other methods of estimating α evaluate the degree of similarity between twin-related intensities. Rees (1982) analysed the behaviour of the R value when calculated as a comparison over potentially twin-related intensities. If twinning is absent, the comparison is effectively between independent intensities, resulting in a random R value, which is 1 for acentric data. In the presence of twinning, the expected R value is $1 - 2\alpha$, providing an equation for extracting α from the R value over twin-related pairs. A related method takes twin-related reflections one pair at a time and evaluates the difference divided by their sum, to give a variable designated H (Yeates, 1988),

$$H \equiv \frac{[I_{\rm obs}(\mathbf{h}_1) - I_{\rm obs}(\mathbf{h}_2)]}{[I_{\rm obs}(\mathbf{h}_1) + I_{\rm obs}(\mathbf{h}_2)]}, \qquad (11.7.7.1)$$

where $I_{\rm obs}(\mathbf{h}_1)$ and $I_{\rm obs}(\mathbf{h}_2)$ are the observed intensities of two twin-related reflections, h_1 and h_2. The value of $|H|$ ranges from 0 to 1 for untwinned (acentric data) and has a uniform probability distribution. A plot of the cumulative distribution of $|H|$ is therefore linear. In the presence of twinning, $|H|$ takes on a restricted range from 0 to $1 - 2\alpha$ but remains uniformly distributed. A plot of $|H|$ therefore provides a relatively simple means of extracting α (Fig. 11.7.7.1). In addition, the expected value and moments for H relate simply to α. The value of $\langle|H|\rangle$ is $(1 - 2\alpha)/2$, while the value of $\langle H^2\rangle$ is $(1 - 2\alpha)^2/3$, providing equations for determining α from measured values of $\langle|H|\rangle$ or $\langle H^2\rangle$.

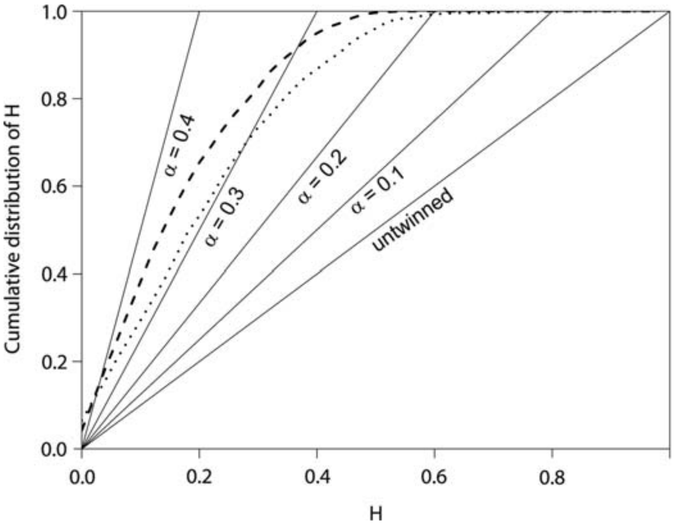


Figure 11.7.7.1
Estimation of the twin fraction α based on the variable H (see text). Two cases of twinning by hemihedry are shown, along with theoretical lines for different values of α. Results from diffraction data on dihaem *c*-type cytochrome DHC2 (PDB code 2czs; Heitmann & Einsle, 2008) are shown as a dashed curve. Results from *Escherichia coli* elongation factor Tu (PDB code 2fx3; Heffron *et al.*, 2006) are shown as a dotted curve. α is typically estimated from the central linear region of the observed curve.

Measurement errors complicate methods for estimating α. In a Britton plot, the ideally sharp transition that indicates the correct value for α becomes more gradual, leading to uncertainty. In methods that use measures of the average difference between twin-related reflections, ignoring the effects of errors causes α to be underestimated. Equations for incorporating the effects of errors have been discussed by Rees (1982), Fisher & Sweet (1980), Dumas *et al.* (1999) and Lunin *et al.* (2007). In general, measurement errors have a relatively small effect on estimates of α when the twin fraction is low, but have a much larger effect when the intensity measurement errors (on a fractional scale) are comparable with the deviation of α from 1/2.

Local or noncrystallographic symmetry (NCS) often presents challenging complications in relation to twinning. If a local symmetry operator nearly coincides with a crystallographic operator of some higher crystal symmetry, this gives rise to higher pseudo-crystallographic symmetry. Especially at lower resolution, reflections related by this operation will have similar intensities, and this can be mistaken for partial twinning. A further complication is that *bona fide* twinning often occurs in conjunction with NCS. One approach for resolving such complications is to analyse intensity statistics as a function of resolution. In cases where NCS is present but twinning is not, the similarity between potentially twin-related reflections typically decreases at higher resolution. In contrast, similarity between twin-related reflections persists out to the maximum resolution if twinning is present.

NCS can also occur in conjunction with perfect twinning. As before, resolution-dependent behaviour can be informative. Tests of overall intensity statistics (*i.e.* tests of perfect twinning) described earlier can appear relatively normal at low resolution where, owing to NCS, the true crystallographic intensities of twin-related reflections may be very similar to each other. If the true crystallographic intensities of twin-related reflections are nearly equal, twinning amounts to an averaging of values that are already nearly equal, which has little effect on intensity statistics. The effects of twinning on intensity statistics may be prominent only at higher resolution, where the true crystallographic intensities of twin-related reflections have diverged.

11.7.8. Higher forms of twinning

Types of twinning beyond hemihedry are possible. When four distinct twin domain orientations are involved, the situation is referred to as tetartohedry. For macromolecular crystals, twinning by tetartohedry is possible when the true symmetry of the crystal is primitive trigonal, *e.g.* $P3$. The crystal lattice has 622 rotational symmetry (order 12), which exceeds the rotational symmetry of $P3$ by a factor of four. Multiple cases have been reported (Rosendal *et al.*, 2004; Barends *et al.*, 2005; Gayathri *et al.*, 2007; Yu *et al.*, 2009). Twinning by pseudo-tetartohedry is possible in symmetries other than $P3$ if the unit-cell geometry obeys fortuitously high symmetry (Anand *et al.*, 2007). In twinning by tetartohedry, four crystallographic intensities contribute to each observed intensity. The effects on intensity statistics can therefore be even more pronounced compared with twinning by hemihedry. This is especially true when all four domain orientations are equally represented (*i.e.* the four underlying twin fractions are all 1/4), a situation referred to as perfect tetartohedry. In that case, the ratio $\langle I^2\rangle/\langle I\rangle^2$ is expected to be 5/4 (Stanley, 1972) and the value of $\langle|L|\rangle$ is expected to be $35/128 \simeq 0.273$ (Table 11.7.6.1). Less extreme values are expected for partial tetartohedry.

Twinning by partial tetartohedry can manifest itself as a crystal that appears to be partially twinned according to multiple distinct twin laws. However, cases of tetartohedry cannot be described properly in terms of a set of estimated hemihedral twin fractions (*e.g.* one for each twin law). Whereas the twin fraction α (and its complement $1 - \alpha$) can be estimated in the case of hemihedry by comparing pairs of twin-related reflections, the four twin fractions (which sum to unity) in the case of tetartohedry must be estimated by performing statistical tests on quadruplets of twin-related reflections. Equations and methods for that purpose are provided by Yeates & Yu (2008).

11.7.9. Other kinds of disorder

Macromolecular crystals can suffer from disorders other than twinning. These are not covered in depth here, but similarities and differences between twinning and a class of problems known as order–disorder (OD) phenomena are worth noting. In OD phenomena, alternative molecular positions or orientations occur over a short range within a crystal, from one layer to another for instance, in a stochastic fashion (Dornberger-Schiff, 1956; Dornberger-Schiff & Grell-Niemann, 1961). When the stochastic variation between layers or between rows is a translation, this is referred to as lattice-translocation disorder (LTD). Several cases of OD have been described and most have been of the LTD type (Wang, Kamtekar *et al.*, 2005; Wang, Rho *et al.*, 2005; Hwang *et al.*, 2006; Tanaka *et al.*, 2008; Zhu *et al.*, 2008; Hare *et al.*, 2009; Tsai *et al.*, 2009), while cases involving alternately oriented layers are known, but less common (Pletnev *et al.*, 2009). Similar to twinning, OD phenomena can complicate structure determination and prevent proper atomic refinement. However, the symptoms and treatments differ from those for twinning. A distinguishing feature is the length scale of the disorder. In OD cases, differently oriented or translated molecules are nearby in the crystal and therefore scatter coherently. The intensity statistics are therefore not shifted in the same way as for twinning. Furthermore, the interference between alternatively positioned molecules in the case of LTD is typically evident by major non-origin peaks in a native Patterson map, often appearing to be impossibly close to the origin. In addition, LTD diffraction patterns often show prominent streaking in all or a subset of reflections. Another situation that often arises in OD cases is the appearance of sterically impossible (*e.g.* interpenetrating) electron density during model building. This can arise when the model accounts for some of the molecules in the unit cell, but others (typically in alternative, mutually exclusive, orientations or positions) have not been accounted for. In contrast, unexplained or interpenetrating electron density is generally not observed in twinning, owing to the absence of coherent scattering between differently oriented twin domains.

11.7.10. Summary

In situations where symmetry permits it, twinning is fairly common in macromolecular crystals. If twinning is present but goes undetected, problems arise in later stages of structure determination. Partial and perfect twinning lead to distinct scenarios, but statistical tests make it possible to uncover both kinds of problem. Local or noncrystallographic symmetry, which very often occurs together with twinning, can complicate the interpretation. Resolution-dependent analysis can help to clarify these cases. Additional challenges occur with more complex types of twinning, such as tetartohedry, which is being recognized with increasing frequency.

The authors thank Michael Sawaya, Regine Herbst-Irmer and George Sheldrick for helpful comments on the manuscript.

References

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C. & Terwilliger, T. C. (2004). *Recent developments in the PHENIX software for automated crystallographic structure determination. J. Synchrotron Rad.* **11**, 53–55.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *PHENIX: building new software for automated crystallographic structure determination. Acta Cryst.* D**58**, 1948–1954.

Anand, K., Schulte, A., Fujinaga, K., Scheffzek, K. & Geyer, M. (2007). *Cyclin box structure of the P-TEFb subunit cyclin T1 derived from a fusion complex with EIAV tat. J. Mol. Biol.* **370**, 826–836.

Barends, T. R. M., de Jong, R. M., van Straaten, K. E., Thunnissen, A.-M. W. H. & Dijkstra, B. W. (2005). *Escherichia coli MltA: MAD phasing and refinement of a tetartohedrally twinned protein crystal structure. Acta Cryst.* D**61**, 613–621.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bruker (2000). *XPREP* Software Reference Manual. Bruker AXS Inc., Madison, Wisconsin, USA.

Dauter, Z. (2003). *Twinned crystals and anomalous phasing. Acta Cryst.* D**59**, 2004–2016.

Dix, I., Sevvana, M., Herbst-Irmer, R. & Sheldrick, G. M. (2007). *Processing and scaling algorithms for small molecules, macromolecules and twins. Acta Cryst.* A**63**, s162.

Dornberger-Schiff, K. (1956). *On order–disorder structures (OD structures). Acta Cryst.* **9**, 593–601.

Dornberger-Schiff, K. & Grell-Niemann, H. (1961). *On the theory of order–disorder (OD) structures. Acta Cryst.* **14**, 167–177.

Duisenberg, A. J. M. (1992). *Indexing in single-crystal diffractometry with an obstinate list of reflections. J. Appl. Cryst.* **25**, 92–96.

Duisenberg, A. J. M., Kroon-Batenburg, L. M. J. & Schreurs, A. M. M. (2003). *An intensity evaluation method: EVAL-14. J. Appl. Cryst.* **36**, 220–229.

Dumas, P., Ennifar, E. & Walter, P. (1999). *Detection and treatment of twinning: an improvement and new results. Acta Cryst.* D**55**, 1179–1187.

Efremov, R., Moukhametzianov, R., Büldt, G. & Gordeliy, V. (2004). *Physical detwinning of hemihedrally twinned hexagonal crystals of bacteriorhodopsin. Biophys. J.* **87**, 3608–3613.

Fisher, R. G. & Sweet, R. M. (1980). *Treatment of diffraction data from crystals twinned by merohedry. Acta Cryst.* A**36**, 755–760.

Friedel, G. (1926). *Leçons de Cristallographie*. Paris: Berger-Levrault.

Gayathri, P., Banerjee, M., Vijayalakshmi, A., Azeez, S., Balaram, H., Balaram, P. & Murthy, M. R. N. (2007). *Structure of triosephosphate isomerase (TIM) from Methanocaldococcus jannaschii. Acta Cryst.* D**63**, 206–220.

Hamdane, D., Lechauve, C., Marden, M. C. & Golinelli-Pimpaneau, B. (2009). *Pseudo-merohedral twinning in monoclinic crystals of wild-type human brain neuroglobin. Acta Cryst.* D**65**, 388–392.

Hare, S., Cherepanov, P. & Wang, J. (2009). *Application of general formulas for the correction of a lattice-translocation defect in crystals of a lentiviral integrase in complex with LEDGF. Acta Cryst.* D**65**, 966–973.

Heffron, S. E., Moeller, R. & Jurnak, F. (2006). *Solving the structure of Escherichia coli elongation factor Tu using a twinned data set. Acta Cryst.* D**62**, 433–438.

Heitmann, D. & Einsle, O. (2008). *Pseudo-merohedral twinning in crystals of the dihaem c-type cytochrome DHC2 from Geobacter sulfurreducens. Acta Cryst.* D**64**, 993–999.

Helliwell, J. R. (2008). *Macromolecular crystal twinning, lattice disorders and multiple crystals. Crystallogr. Rev.* **14**, 189–250.

Herbst-Irmer, R. & Sheldrick, G. M. (1998). *Refinement of twinned structures with SHELXL97. Acta Cryst.* B**54**, 443–449.

Herbst-Irmer, R. & Sheldrick, G. M. (2002). *Refinement of obverse/reverse twins. Acta Cryst.* B**58**, 477–481.

Hwang, W. C., Lin, Y., Santelli, E., Sui, J., Jaroszewski, L., Stec, B., Farzan, M., Marasco, W. A. & Liddington, R. C. (2006). *Structural basis of neutralization by a human anti-severe acute respiratory syndrome spike protein antibody, 80R. J. Biol. Chem.* **281**, 34610–34616.

Ko, T.-P., Kuznetsov, Y. G., Malkin, A. J., Day, J. & McPherson, A. (2001). *X-ray diffraction and atomic force microscopy analysis of twinned crystals: rhombohedral canavalin. Acta Cryst.* D**57**, 829–839.

Lunin, V. Y., Lunina, N. L. & Baumstark, M. W. (2007). *Estimates of the twinning fraction for macromolecular crystals using statistical models accounting for experimental errors. Acta Cryst.* D**63**, 1129–1138.

MacRae, I. J. & Doudna, J. A. (2007). *An unusual case of pseudo-merohedral twinning in orthorhombic crystals of Dicer. Acta Cryst.* D**63**, 993–999.

Padilla, J. E. & Yeates, T. O. (2003). *A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. Acta Cryst.* D**59**, 1124–1130.

Parsons, S. (2003). *Introduction to twinning. Acta Cryst.* D**59**, 1995–2003.

Pletnev, S., Morozova, K. S., Verkhusha, V. V. & Dauter, Z. (2009). *Rotational order–disorder structure of fluorescent protein FP480. Acta Cryst.* D**65**, 906–912.

Redinbo, M. R. & Yeates, T. O. (1993). *Structure determination of plastocyanin from a specimen with a hemihedral twinning fraction of one-half. Acta Cryst.* D**49**, 375–380.

Rees, D. C. (1980). *The influence of twinning by merohedry on intensity statistics. Acta Cryst.* A**36**, 578–581.

Rees, D. C. (1982). *A general theory of X-ray intensity statistics for twins by merohedry. Acta Cryst.* A**38**, 201–207.

Rosendal, K. R., Sinning, I. & Wild, K. (2004). *Crystallization of the crenarchaeal SRP core. Acta Cryst.* D**60**, 140–143.

Rudolph, M. G., Wingren, C., Crowley, M. P., Chien, Y. & Wilson, I. A. (2004). *Combined pseudo-merohedral twinning, non-crystallographic symmetry and pseudo-translation in a monoclinic crystal form of the T-cell ligand T10. Acta Cryst.* D**60**, 656–664.

Sheldrick, G. M. (1997). *SHELX*. University of Göttingen, Germany.

Sheldrick, G. M. (2002). *TWINABS*. University of Göttingen, Germany.

Sheldrick, G. M. (2004). *CELL_NOW*. University of Göttingen, Germany.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Sparks, R. A. (1999). *GEMINI*. Bruker AXS, Madison, Wisconsin, USA.

Stanley, E. (1972). *The identification of twins from intensity statistics. J. Appl. Cryst.* **5**, 191–194.

Tanaka, S., Kerfeld, C. A., Sawaya, M. R., Cai, F., Heinhorst, S., Cannon, G. C. & Yeates, T. O. (2008). *Atomic-level models of the bacterial carboxysome shell. Science*, **319**, 1083–1086.

Tsai, Y., Sawaya, M. R. & Yeates, T. O. (2009). *Analysis of lattice-translocation disorder in the layered hexagonal structure of carboxysome shell protein CsoS1C. Acta Cryst.* D**65**, 980–988.

Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Correction of X-ray intensities from single crystals containing lattice-translocation defects. Acta Cryst.* D**61**, 67–74.

Wang, J., Rho, S.-H., Park, H. H. & Eom, S. H. (2005). *Correction of X-ray intensities from an HslV-HslU co-crystal containing lattice-translocation defects. Acta Cryst.* D**61**, 932–941.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.

Yeates, T. O. (1988). *Simple statistics for intensity data from twinned specimens. Acta Cryst.* A**44**, 142–144.

Yeates, T. O. (1997). *Detecting and overcoming crystal twinning. Methods Enzymol.* **276**, 344–358.

Yeates, T. O. & Fam, B. C. (1999). *Protein crystals and their evil twins. Structure*, **7**, R25–R29.

Yeates, T. O. & Yu, F. (2008). *Equations for determining tetartohedral twin fractions. Acta Cryst.* D**64**, 1158–1164.

Yu, F., Song, A., Xu, C., Sun, L., Li, J., Tang, L., Yu, M., Yeates, T. O., Hu, H. & He, J. (2009). *Determining the DUF55-domain structure of human thymocyte nuclear protein 1 from crystals partially twinned by tetartohedry. Acta Cryst.* D**65**, 212–219.

Zhu, X., Xu, X. & Wilson, I. A. (2008). *Structure determination of the 1918 H1N1 neuraminidase from a crystal with lattice-translocation defects. Acta Cryst.* D**64**, 843–850.

Zwart, P. H., Grosse-Kunsteleve, R. W. & Adams, P. D. (2005*a*). *Characterization of X-ray data sets. CCP4 Newslett.* **42**, contribution 10.

Zwart, P. H., Grosse-Kunsteleve, R. W. & Adams, P. D. (2005*b*). *Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. CCP4 Newslett.* **43**, contribution 7.

International Tables for Crystallography (2012). Vol. F, Chapter 12.1, pp. 317–326.

PART 12. ISOMORPHOUS REPLACEMENT

Chapter 12.1. The preparation of heavy-atom derivatives of protein crystals for use in multiple isomorphous replacement and anomalous scattering

D. Carvin, S. A. Islam, M. J. E. Sternberg and T. L. Blundell

12.1.1. Introduction

The traditional method of multiple isomorphous replacement (MIR) was introduced by Perutz and co-workers in 1954 (Green *et al.*, 1954) and is often enhanced by anomalous scattering (MIRAS) [see Blundell & Johnson (1976) for a review]. The method remains popular for solution of the phase problem in the absence of the structure of a close homologue, although the use of multiple anomalous dispersion is likely to increase in the coming years (Hendrickson, 1985).

Protein crystals comprise an open lattice of protein molecules with solvent occupying the channels and spaces which normally comprise between 30 and 80% of the crystal volume. The preparation of a useful derivative requires the binding of a heavy atom to a specific position, usually on the protein surface, for example by the displacement of a lighter solvent molecule or an ion, without distorting the protein or crystal lattice.

Ideally, rational selection of suitable heavy-atom reagents requires a comprehensive knowledge and understanding of the crystalline structure of the protein. Normally, this information is unavailable since it is the objective of the crystal structure analysis! Nevertheless, the sequence and mechanism of action may suggest which heavy-atom reagents might be employed. There are reports in the literature of many attempts to make synthetic analogues of specific amino acids, by substituting selenium for sulfur residues in a chemically synthesized polypeptide or by removing an amino-terminal residue by the Edman technique and replacing it with an amino acid modified by a heavy atom [see Blundell & Johnson (1976) for a review]. Alternatively, analogues of the substrate of an enzyme or carrier protein can sometimes be modified with a heavy atom; however, this will disturb the active site, which is usually the region of greatest interest to the structural biologist. Such methods have not proved very useful and will not be described further here. Most proteins studied now are recombinant; site-directed mutagenesis can replace methionines in the sequence, which occur on average once every fifty residues, by selenomethionines (Hendrickson *et al.*, 1990) or more recently by telluromethionines (Budisa *et al.*, 1997). Such approaches have revolutionized macromolecular crystallography through the use of anomalous-dispersion techniques, but have yet to provide a very efficient method of introducing atoms heavier than selenium into proteins.

Thus, the vast majority of successful heavy-atom derivatives employed in crystallographic analyses are obtained on a trial-and-error basis. In earlier studies, the protein was often covalently modified, purified and characterized before crystallization. There are some useful covalent modifications, for example, the reaction of mercury with the sulfhydryl groups of cysteinyl side chains and the iodination of tyrosyl side chains. The replacement of a metal-ion cofactor, such as calcium or zinc, can also give a useful derivative. However, pre-reaction of the protein often gives rise to conformational changes in the protein, and crystallization frequently occurs in a different or non-isomorphous form.

Most heavy-atom derivatives are produced by direct soaking of the crystals in a solution of the heavy-atom compound. With this approach, heavy-atom substitution patterns tend to be complex, with sites frequently only partially occupied. The specificity is often determined by entropic factors. Thus, sites between molecules in the crystal lattice, or between several different side chains brought together by the tertiary structure, may bind the metal ion, even if the side chains individually do not have strong affinity for the metal. Chelation is entropically driven, and bonds may form with unusual protein ligands, a major factor causing lack of specificity.

Blake (1968) reviewed the data available for heavy-atom binding to proteins and suggested some generalizations. These were extended in a comprehensive review of protein heavy-atom derivatives (Blundell & Johnson, 1976; Blundell & Jenkins, 1977) which analysed the dependence of reactivity on protein side chain identity, nature of the reagent, pH, concentration, buffer *etc*. Over the past two decades, there have been discussions of the binding of some particular metal ions, but there have been no comprehensive analyses. Furthermore, protein–heavy-atom interactions have not often been fully described in publications of protein crystallographic analyses, and, in any case, the information has not been available in a format that could be used for systematic computer-based analysis.

We have now collected, either from the literature or directly from protein crystallographers, information on the preparation and characterization of heavy-atom derivatives of protein crystals. We have defined heavy atoms as those with an atomic weight greater than that of rubidium. We have assembled the information in the form of a data bank (Carvin *et al.*, 1991; Islam *et al.*, 1998) in which the coordinate data for the heavy-atom positions are compatible with the crystallographic data in the Protein Data Bank (Bernstein *et al.*, 1977). The data bank contains a wealth of information and provides the basis for further, more detailed analyses of heavy-atom binding to proteins. The information can be directly accessed and should be useful to protein crystallographers seeking to improve their success in preparing heavy-atom derivatives for isomorphous replacement and anomalous dispersion.

In this chapter we provide an introduction to the data bank and we review strategies that can be adopted in the preparation of heavy-atom derivatives of protein crystals for use in MIRAS.

12.1.2. Heavy-atom data bank

The heavy-atom data bank (HAD) is a computer-based archival file system that contains experimental and derived information from successful multiple isomorphous replacement analyses in the determination of protein crystal structures. HAD is available

Copyright © 2012 International Union of Crystallography

at http://www.sbg.bio.ic.ac.uk/had/. The data bank makes available information which is otherwise only accessible in a widely distributed and fragmented form throughout the scientific literature or even unpublished in laboratory files.

The data bank contains information on heavy-atom derivatives for 969 protein crystals, 600 of which are deposited in the Protein Data Bank (PDB). A further 200 proteins are being processed at present. It contains information on the physical and chemical characteristics of each chemical compound that has proved successful in past protein crystallographic analyses: this includes the IUPAC name, trivial name, molecular formula, oxidation state, solution chemistry and stereochemistry. Experimental details of the preparation of the heavy-atom derivatives include the source of the protein, concentration of the heavy-atom solution, pH values, soak times and details of the buffer compositions used in the experiments; 3164 different experimental conditions are recorded. The atomic coordinates are given in the same format as the PDB coordinates for the 5500 binding sites of the heavy atoms. A statistical analysis is included for each of the 456 heavy-atom reagents; this includes range of pH values and a summary of the amino acids involved at the binding sites. For metalloproteins, it gives details of the type, number, geometry of coordination and function of the native metal(s) present. This is followed by a description of the procedure for native-metal substitution and details of the coordination of the substituted heavy atom. It also includes an extensive bibliography and references to other relevant web sites.

12.1.3. Properties of heavy-atom compounds and their complexes

Potential ligands for heavy-atom reagents may be derived from the functional group(s) of reactive amino-acid side chains, from the buffer and from salting out/in agents. We must first consider factors that will influence the formation of such complexes in the environment of a protein crystal.

12.1.3.1. Stability

Ligands may be classified as either hard or soft. Hard ligands tend to be electronegative and interact electrostatically, with little delocalization of electron density. Water molecules, glutamates, aspartates, terminal carboxylates, and hydroxyl groups of serine and threonine from the protein, as well as acetate and citrate ions from the buffer, fall into this category. Conversely, soft ligands are polarizable and tend to form covalent bonds. Typical examples include the anions Cl^-, Br^-, I^-, S^{2-}, CN^-, imidazole, methionine, cysteine, cystine and histidine from the protein. Ligands can be listed in series of increasing hardness:

$$I^- > Br^- > Cl^- \gg F^- \gg H_2O,$$
$$RS^- > R_2S \gg NH_3 > H_2O,$$
$$CN^- > RNH_2 > Cl^- > CO_2^- > \text{—OH of alcohol.}$$

The metal components of the reagents may be classified as hard (class A) or soft (class B) in a similar way. Class A metals include the alkali metals, the alkaline earth metals, the lanthanide and actinide series, and the first-row transition metals from group III to group VA. Many of these metal ions have an inert-gas structure in which the electrons are held very strongly and tend to be non-polarizable. Metal ions in this class tend to interact with hard ligands, including the acetate, citrate and phosphate buffer components of mother liquor systems. On the other hand, class B metals have a preference for binding soft ligands. This group includes most members of the second and third row of the transition series (*e.g.* Ag, Cd, Pt, Au, Hg), which form cations such as $Pt(NH_3)_4^{2+}$ or anions such as $Au(CN)_2^-$, $PtCl_4^{2-}$ and HgI_4^{2-}. The easily polarizable *d* electrons allow formation of covalent bonds with methionine, cysteine and imidazole, so displacing the ligands of the complexes. In the middle and towards the end of the first transition-metal series, the ions have properties intermediate between class A and B metals. Class B character increases in the series:

$$Fe^{2+} < Co^{2+} < Ni^{2+} < Cu^{2+} < Zn^{2+}.$$

Thus, zinc binds to the polarizable sulfur of cysteine and imidazole of histidine as well as to carboxylates and water molecules. Tl^+ and Pb^{2+}, which each have an inert pair of electrons in their outer shell, are stable cations and prefer carboxylate rather than sulfur ligands or imidazole.

12.1.3.2. Lability

The rates at which ligands enter and leave a metal complex are important in the formation of heavy-atom derivatives, especially the covalent complexes of mercury, gold and platinum. The rate-determining step in unimolecular S_{N1} reactions is the expulsion of the leaving ligand from the metal complexes, which often proceeds relatively slowly. The intermediate complex, once formed, reacts with the entering ligand almost instantly. For S_{N1} reactions, the rate is directly proportional to the intermediate complex concentration but independent of the ligand concentration. The bimolecular S_{N2} mechanism involves attack by the ligand on the metal complex to form an intermediate complex, which then ejects the displaced ligand. The rate of reaction is proportional to the concentration of the initial species and the concentration of the nucleophile. S_{N2} reaction rates are dependent on the nature of the leaving group and the attacking nucleophile in the following ways:

$$\text{Relative rates of attack: } RS^- > I^- > Br^- > NH_3 > Cl^- > RO^-;$$
$$\text{Rate of leaving group: } H_2O > Cl^- > NO_2^- > CN^-.$$

Sulfur ligands are good nucleophiles but poor leaving groups. They form thermodynamically stable complexes. The rate of leaving is influenced by the *trans* effect in square-planar complexes of Au(III) and Pt(II). Thus groups in square-planar complexes *trans* to NH_3 are difficult to displace. This has implications for attempts to make derivatives of proteins in ammonium sulfate, where ligands may be replaced by NH_3.

Rates of reaction depend not only upon which ligands are present in a heavy-atom complex but also on the character of the metal. For example, $PtCl_4^{2-}$, $AuCl_4^-$ and $PdCl_4^-$ have similar square-planar geometries (Petsko *et al.*, 1978), but the rates of substitution vary:

$$PdCl_4^- > PtCl_4^{2-} > AuCl_4^-.$$

Thus, if the reaction between the protein and a palladium or platinum complex is proceeding too fast, a gold derivative might be investigated.

12.1.3.3. Oxidation state of metal ions in protein crystals

In the environment of a living cell, the following oxidation states tend to be stable:

Os(II), Ru(II), Ir(III), Rh(III), Pt(II), Pd(II), Au(I).

12.1.3.4. Effect of pH

Although the pK_a of an individual amino acid in solution is generally defined within narrow limits, environmental and steric

Table 12.1.3.1

Useful pH ranges of some heavy-atom reagents derived from the heavy-atom data bank

No. of entries	Minimum	Average	Maximum	Compound
159	3.0	6.7	9.1	Potassium tetrachloroplatinum(II)
63	4.2	6.6	9.0	Potassium dicyanoaurate(I)
53	4.2	6.9	9.5	Mercury(II) chloride
59	2.8	6.7	9.0	Mercury(II) acetate
52	4.7	6.7	9.3	4-(Chloromercurio)benezenesulfonic acid
57	2.0	6.5	9.3	Potassium tetraiodomercurate(II)
36	5.4	6.7	8.5	Ethylmercurythiosalicylate (EMTS)
46	4.0	6.0	8.0	Potassium pentafluorooxyuranate(VI)
2	8.2	8.4	8.5	Barium(II) chloride
22	4.0	6.2	8.1	Lead(II) acetate
13	4.5	6.6	7.5	Lead(II) nitrate
1	6.5	6.5	6.5	Strontium(II) acetate
3	6.3	6.8	7.5	Thallium(I) acetate
2	5.9	6.6	7.2	Thallium(III) chloride
5	5.0	5.8	6.8	Gadolinium(III) chloride
9	4.9	6.7	7.5	Samarium(III) nitrate
7	4.9	6.6	8.7	Neodymium(III) chloride
64	4.1	6.3	8.6	Uranium(VI) oxyacetate

factors give rise to a wide range of values in proteins. Thus, the hydrogen-ion concentration influences the thermodynamic and kinetic stability of potential complexes. Protons compete with heavy-atom ions for the available binding site(s) on the protein. For example, below pH 3.5, cations bind less well to aspartic and glutamic acids due to the protonation of the carboxylate groups.

The nucleophilicity of histidine increases when it loses its proton, and thus its positive charge changes from around pH 6.0 to 7.0. Similarly, the nucleophilicity of cysteine increases dramatically when the thiolate ion is formed at pH $\sim$ 8.0. The thiolate ion is a stronger nucleophile than the thioether group of methionine, but when it becomes protonated it is considerably less effective. The nucleophilicity of the attacking groups varies in the order

$$RS^- > R_2S > RSH.$$

Thus the number and occupancy of sites can be manipulated by varying the pH, often after cross-linking the crystals to stabilize them.

Extremes in pH can give rise to considerable difficulties in establishing suitable derivatives, as hydrogen and hydroxyl ions compete with the metal ion/complex for the protein and with the protein for the metal ion/complex. At extremely high pH values metals in solution tend to form insoluble hydroxides. The ranges of pH values that are useful for metal ions are given in Table 12.1.3.1.

Varying the reactivity of amino-acid side chains by manipulation of the pH can enable the same heavy-atom ion/complex to bind at different sites, thus producing more than one derivative useful for phase determination.

12.1.3.5. Effect of precipitants and buffers on heavy-atom binding

Components present in the heavy-atom solution can have a profound effect on protein–heavy-atom interactions. The salting in/out agent (precipitant) and buffer are the principal sources of alternative ligands for the heavy-atom reagents, while protons compete with the heavy-atom ion/complex for the reactive amino-acid side chains.

Ammonium sulfate is the most successful precipitant in protein crystallization experiments (Gilliland *et al.*, 1994). However, its continued presence in the mother liquor can cause problems by interfering with protein–heavy-atom interactions. At high hydrogen-ion concentrations, the NH_3 group is protonated (*i.e.* NH_4^+), but as the pH rises the proton is lost, typically around pH 6.0–7.0, enabling the group to compete with the protein for the heavy-atom reagent by an S_{N2} reaction.

The nucleophilic strength of potential ligands follows the order

$$NH_3 > Cl^- > H_2O.$$

The anionic complex $PtCl_4^{2-}$ is present in excess ammonia at pH > 7.0 and it will react:

$$PtCl_4^{2-} \rightarrow cis\text{-}PtCl_2(NH_3)_2 \rightarrow Pt(NH_3)_4^{2+}.$$

The resultant cationic complex is less susceptible to reaction due to the *trans* effect of NH_3. Pd, Au, Ag and Hg complexes react in a similar way. Decreasing the pH of the solution reduces the amount of free ammonia available through protonation (Sigler & Blow, 1965). Such a technique may give rise to other problems (*e.g.* cracked crystal, decreased nucleophilicity of the protein ligands).

Changing the precipitant to sodium/potassium phosphate or magnesium sulfate may alleviate the situation, but it may also present other problems. For instance, PO_4^{3-} displaces Cl^- from $PtCl_4^{2-}$, thus increasing the negative charge. Both PO_4^{3-} and SO_4^{2-} form insoluble complexes with class A metals (*e.g.* lanthanide and uranyl cations) (Petsko *et al.*, 1978). Both acetate and citrate form complexes with class A metals, but citrate, a chelating ion, binds more strongly. Tris buffer is probably preferable; it binds many cations, but the complexes formed tend to be relatively unstable.

12.1.3.6. Solubility of heavy-atom compounds

The solubility of a heavy-atom compound will depend upon the precipitant, buffer and pH. Typically, the component present in the highest concentration is the precipitant, either as salts (*e.g.* ammonium sulfate) or as an organic-based reagent (*e.g.* ethanol, MPD, PEG). Heavy-atom compounds that are essentially covalent and organic in character will be more soluble in ethanol, MPD, PEGs and other organic precipitants.

Although the solubility of tetrakis(acetoxymercurio)methane (TAMM) is higher than most multiple-heavy-atom compounds in aqueous solutions, the presence of glycylglycine or charged mercaptans, such as cysteamine or penicillamine, can increase solubility further (Lipka *et al.*, 1976). The ratio of TAMM to solubilization agent (*e.g.* glycylglycine) is typically 1:10. Even so, the final solubility of TAMM depends on the concentration of competing anions (*e.g.* chloride) (O'Halloran *et al.*, 1987).

Many organometallic compounds are relatively insoluble in aqueous solutions, but their solubility may be increased by predissolving in an aprotic solvent such as acetonitrile.

Iodine and several inorganic iodide salts are insoluble in aqueous solutions. This can be rectified by dissolving the heavy-atom compounds in an aqueous solution of KI.

12.1.3.7. Effect of concentration, time of soak and temperature on heavy-atom binding

Most heavy-atom derivatives are prepared by diffusing or dialysing the compound into the crystals. Concentrations have typically ranged from 0.1–100.0 m*M*. Occasionally, concentrations as low as 0.001 m*M* have been employed to maintain crystal integrity. Low concentrations favour sites where the interactions between the heavy atom and the protein ligands are strongest. Decreasing the number of non-specific interactions minimizes the amount of heavy-atom reagent in the lattice. The latter absorbs X-rays without contributing to the diffraction pattern except at low angles. Increasing the concentration may give rise to other binding site(s). Usually, the higher the concentration employed, the shorter the soak time required for equivalent substitution. Short soak times at high concentrations tend to denature the crystals more often than long soaks at low concentrations. At very high concentrations (*i.e.* > 100 m*M*), the heavy-atom compound perturbs the protein crystal–mother liquor equilibrium by withdrawing water molecules from the hydration shell around the periphery of the crystal. Disorder of the crystals can sometimes be avoided by the application of a cross-linking reagent (*e.g.* glutaraldehyde). The optimal concentration is the lowest concentration that consistently reproduces intensity differences in the diffraction pattern of 15–25% without cracking and disordering the crystals.

Length of soak may be important. The heavy-atom data bank shows that, typically, soak times range from one day to one week. Useful derivatives have been prepared with a soak time of an hour to over a year. If no binding is apparent after several days, extending the soak time to over a week may produce some binding, but this is rare. Soaks of 24 hours for simple inorganic salts and up to one week for other types of heavy-atom compounds will normally suffice when screening for binding. The concentration of the heavy-metal compound that can be achieved will depend on its solubility in the crystal stabilization solution. Normally, the longer the soak, the greater the occupancy. Exceptions can arise due to undesirable chemical reactions between components present in the derivatization solution.

For covalent-bond formation, the length of soak and the concentration can be short (*e.g.* 1 h, 0.01 m*M*). This is especially true for mercury derivatives of proteins that have reactive sulfhydryls (Ringe *et al.*, 1983).

Variations in the temperature can also alter the rate of reaction. The UO_2 acetate derivative of rhombohedral insulin binds twenty times more slowly at 4 °C than at ambient temperature (Blundell, 1968). A lower temperature allows greater control over the rate of substitution. Conversely, heavy-atom derivatives that do not appear to bind may do so upon elevation of the temperature.

12.1.4. Amino acids as ligands

The reactivity of the heavy-atom reagent will also depend on the state of the amino-acid residues in the protein.

The thiolate anion of *cysteine*, a potent nucleophile, reacts almost irreversibly with mercuric complexes or organomercurials. It also acts as a fast-entering attacking group in S_{N2} ligand substitution reactions with other class B metals (*e.g.* Ag, Ir, Rh, Pt, Pd, Au), forming stable complexes. Below pH 6, the thiolate anion becomes protonated. As covalent reactions are less sensitive to hydrogen-ion concentration than ligand substitution reactions, cysteines still bind rapidly with mercurials, but there is negligible reaction with other class B metals (Petsko *et al.*, 1978).

Cystines are very weakly reactive in ligand substitution reactions. However, $PtCl_4^{2-}$ binds to disulfides in some proteins with displacement of a chloride ion (Lipscomb *et al.*, 1970; Sigler *et al.*, 1968). Mercurials rarely insert spontaneously into disulfide linkages. However, substitution of mercury can be achieved either by the prior application of a reducing agent such as dithiothreitol (Ely *et al.*, 1973; Sperling *et al.*, 1969), or by direct application of reducing mercurous ions (Sperling & Steinberg, 1974).

The non-ionizable thioether group of *methionine* is unreactive towards mercurials, but the lone pair of electrons on sulfur allows nucleophilic S_{N2} ligand substitution. Methionine will displace Cl, I, Br and NO_2 ligands from platinum complexes to form a stable bond. The reaction of methionine with platinum compounds is not pH sensitive within the normal range. The residue may become unreactive through oxidation, first to the sulfoxide and then to the sulfone; only the sulfoxide can be reduced readily by thiols or other reducing agents.

Below pH 6, *histidine* exists mainly as an imidazolium cation. Although this is not reactive as a nucleophile, it can interact electrostatically with anionic complexes. At pH 7 and above, the unprotonated imidazole is a good nucleophile, being able to displace Cl, Br, I and NO_2 ligands from platinum, silver, mercury and gold complexes. Electrophilic substitution of iodine in the imidazole ring is feasible, but the conditions are severe and it has not proved very useful in preparing derivatives.

At pH < 8.5, the ε-amino group of *lysine* is protonated, allowing it to form weak electrostatic interactions with anionic heavy-atom complexes, but not to participate in S_{N2} substitution reactions. Above pH 9, the free amino group can displace Cl but not Br, I or NO_2 ligands from platinum and gold complexes. The pK_a of the guanidinium group of *arginine* is very high (>12 in proteins), so it interacts electrostatically as a cation with heavy-atom anionic complexes.

The indole ring of *tryptophan* is relatively inert to electrophilic substitution by iodine, but the ring nitrogen can be mercurated (Tsernoglou & Petsko, 1976). The reaction is not pH dependent, but there should be no competing nucleophiles in the mother liquor. Tryptophan does not usually participate as a ligand in substitution of heavy-atom complexes.

The phenolate oxygen anion of *tyrosine* is a good nucleophile and has the potential to bind a substantial number of heavy-atom complexes *via* S_{N2} ligand substitution reactions. However, it has a very high pK_a value of 10.5. Below pH 10, the protonated oxygen predominates, making electrophilic aromatic substitution by iodine the principal reaction.

Aspartic and glutamic acids have side-chain pK_a values in the range 3 to 4. At low pH, they will be protonated and unreactive. Above pH 5, the side chains will be anionic, making them good

Table 12.1.5.1
The 23 most commonly used heavy-atom reagents

The first column gives the number of times the reagent has been used in the analyses included in the heavy-atom data bank.

No.	Compound
287	Potassium tetrachloroplatinum(II)
111	Potassium dicyanoaurate(I)
103	Uranyl acetate
101	Mercury(II) acetate
98	Mercury(II) chloride
85	Ethylmercurythiosalicylate (EMTS)
82	Potassium tetraiodomercurate(II)
81	*para*-Chloromercuriobenzenesulfonate (PCMBS)
75	Trimethyllead(IV) acetate
73	Potassium pentafluorooxyuranate(VI)
73	Phosphatotris(ethylmercury)
61	Potassium tetranitritoplatinum(II)
60	Uranyl nitrate
58	Potassium tetracyanoplatinum(II)
57	Dichlorodiammineplatinum(II)
51	Potassium hexachloroplatinum(IV)
51	Methylmercury chloride
44	Potassium tetrachloroaurate(III)
42	*para*-Chloromercurybenzoate (PCMB)
39	Lead(II) acetate

ligands for class A cations such as uranyl and rare earths. *Glutamine and asparagine* take part in metal coordination but rarely bind strongly enough to metal ligands on their own.

Hydroxyl groups of *serines and threonines* are fully protonated at normal pH values and are consequently not reactive nucleophiles. Abnormally reactive serines, usually at the active site as in serine proteases and β-lactamases, can react with heavy-atom reagents to give useful derivatives.

12.1.5. Protein chemistry of heavy-atom reagents

The heavy-atom data bank (Islam *et al.*, 1998) can be used to analyse the most commonly used heavy-atom reagents: these are given in Table 12.1.5.1. This shows that platinum, gold, mercury and uranyl have provided the most useful reagents.

The heavy-atom data bank can be used as a source of information about the reactivity of proteins to different heavy-atom reagents. This provides the basis for the following analysis.

12.1.5.1. Hard cations

Uranyl-ion complexes have proved the most popular A-group metal reagents for preparing heavy-atom derivatives of protein crystals (see Table 12.1.5.1). UO_2^{2+} is a linear, covalent group based on U(VI), the most stable oxidation state of uranium. Table 12.1.5.2 lists the most commonly used uranyl derivatives. Uranyl compounds may show 2 + 4, 2 + 5, or 2 + 6 coordination, with ligands lying in or near a plane normal to the $O{=}U{=}O^{2+}$ axis. These equatorial ligands may be neutral (*e.g.* H_2O) or anionic (*e.g.* NO_3^-, CH_3COO^-, oxalate^{2-}, F^-, Cl^1 or O_2^-); the nitrate and acetate are bidentate ligands. An example is given in Fig. 12.1.5.1. Anionic complexes, such as $UO_2F_5^{3-}$, have been found near negatively charged amino-acid residues (*e.g.* Glu and Asp), suggesting that the equatorial ligands have been displaced. At low pH, uranyl groups have been located near the hydroxyl groups of threonine and serine residues.

The fifteen lanthanides have similar chemical properties and are generally used as nitrates, acetates or chlorides (Blundell & Johnson, 1976; Carvin, 1986). The lanthanide contraction, a steady decrease in size with increasing atomic number, allows selection of an ion with a radius that will give high occupancy and isomorphism. Gadolinium and samarium salts have the added advantage that the number of anomalous electrons is high.

Table 12.1.5.2
The five most popular uranium derivatives

The first column gives the number of times the reagent has been used in the analyses included in the heavy-atom data bank.

No.	Compound
103	Uranyl acetate
73	Potassium pentafluorodioxyuranate(VI)
60	Uranyl nitrate
8	Uranium(VI) oxysulfate
4	Sodium triacetatedioxyuranate(VI)

Lanthanide ions have greater selectivity than the uranyl ion, which often forms clusters on the protein surface. Uranyl complexes and lanthanide ions are not very soluble above pH 7 and pH 9, respectively, due to the formation of hydroxides. Phosphate buffers should be avoided since they will compete for the heavy atom, often giving insoluble phosphates. In the presence of citrate, samarium is chelated and, since the citrate is difficult to replace, reaction may be inhibited. However, exchanging the buffer for Tris or acetate may enable a useful derivative to be obtained.

12.1.5.2. Thallium and lead ions

Thallium and lead can provide useful derivatives, especially in their lower oxidation states, Tl(I) and Pb(II), when they resemble class A metals. Owing to the non-group valence and presence of an inert pair of electrons, the ionic radii of Tl^+ (1.44 Å) and Pb^{2+} (1.21 Å) are greater than most class A metals. Thallous and plumbous cations prefer carboxylate rather than imidazole or sulfur ligands, although Pb^{2+} occasionally manifests its intermediate character by interacting with imidazole groups. Thallic (Tl^{3+}) and plumbic (Pb^{4+}) ions are similar to class B metals, showing preferential binding to soft ligands, but they are easily reduced in protein solutions.

12.1.5.3. B-metal reagents

The most useful members of the B-metal group, platinum, gold and mercury, give rise to an extensive range of heavy-atom

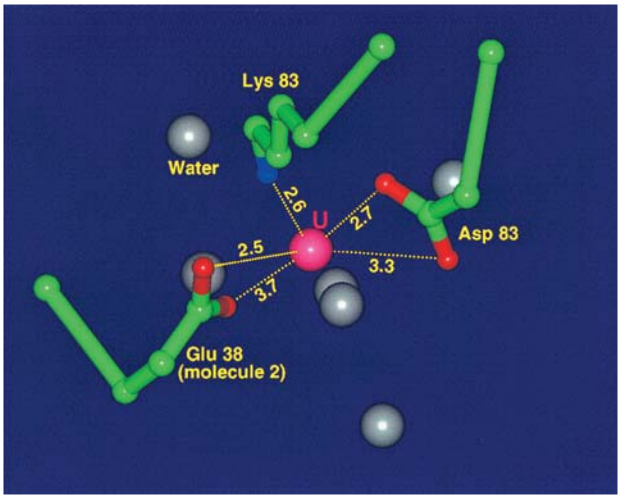


Figure 12.1.5.1
The binding site for uranyl ions in cytochrome *b*5 (oxidized: 3B5C). The positions of the ligands in the parent crystals are shown; these probably move in the complex.

Table 12.1.5.3
The five most popular mercury derivatives

The first column gives the number of times the reagent has been used in the analyses included in the heavy-atom data bank.

No.	Compound
101	Mercury(II) acetate
98	Mercury(II) chloride
85	Ethylmercurythiosalicylate (EMTS)
82	Potassium tetraiodomercurate(II)
81	*para*-Chloromercuriobenzenesulfonate (PCMBS)

compounds which form covalent, electrostatic and van der Waals complexes with proteins. Some compounds can bind to the protein molecule in different ways; for example, $PtCl_4^{2-}$ can bind either covalently to the thioether group of methionine or electrostatically with positively charged residues.

Mercury compounds have proved very successful for preparing heavy-atom derivatives of protein crystals (Table 12.1.5.1), mainly due to the ease of formation of covalent bonds with cysteine residues. An example is given in Fig. 12.1.5.2 in which mercuric chloride has been used to replace zinc in thermolysin. Hg^{2+} complexes are commonly two-coordinate linear and four-coordinate tetrahedral. The most popular mercury reagents are given in Table 12.1.5.3. The covalent character in $Hg{-}L$ bonds, especially in the two-coordinate complexes, can cause solubility problems in aqueous solutions. However, an excess of an alkali metal salt (*e.g.* $HgI_2 + 2KI \rightarrow K_2HgI_4$) will often convert the compound to a more soluble anionic complex of the type HgX_4^{2-}, where X = Cl^-, Br^-, I^-, SCN^-, NCS^-, CN^-, SO_4^{2-}, oxalate^{2-}, NO_3^- or NO_2^-. In the presence of ammonium salts at high pH values, the cationic tetraammine complex, $Hg(NH_3)_4^{2+}$, tends to form. Variation in the charge on the aromatic groups of organomercurials can give rise to different substitution patterns.

Silver, used as the nitrate, tends to interact with cysteine or histidine (see Fig. 12.1.5.3). In the presence of ammonium sulfate, it probably reacts as the ammonia complex, $Ag(NH_3)_4^+$. Silver ions are less polarizing and less reactive than Hg^{2+} ions; thus they give similar derivatives but often with less disorder, as in glucagon (Sasaki *et al.*, 1975). Where the metal ion displaces a proton, Ag^+ will need to react at a higher pH than Hg^{2+}.

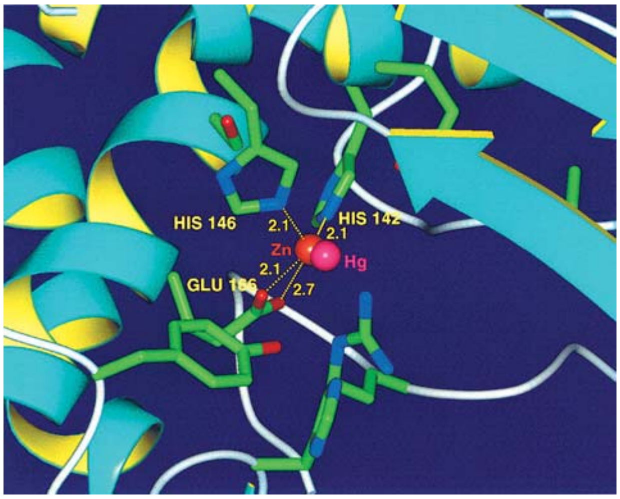


Figure 12.1.5.2
Mercuric ions replace zinc in thermolysin (3TLN). The mercuric ion is shown superposed on the parent crystal structure; notice that the mercuric ion is slightly displaced from the zinc position due to its larger ionic radius.

Table 12.1.5.4
The five most popular platinum derivatives

The first column gives the number of times the reagent has been used the analyses included in the heavy-atom data bank.

No.	Compound
287	Potassium tetrachloroplatinum(II)
61	Potassium tetranitritoplatinum(II)
58	Potassium tetracyanoplatinum(II)
57	Dichlorodiammineplatinum(II)
51	Potassium hexachloroplatinum(IV)

The class B metals *palladium*, *platinum* and *gold* form stable covalent complexes with soft ligands, such as chloride, bromide, iodide, ammonia, imidazole and sulfur groups. The stereochemistry of their complexes depends on the number of d electrons present. For instance, the d^{10} ion of Au(I) gives a linear coordination of two [*e.g.* $Au(CN)_2^-$], whereas d^8 ions of Pd(II), Pt(II) and Au(III) are predominantly square planar, giving cationic [*e.g.* $Pt(NH_3)_4^{2+}$], anionic [*e.g.* $Au(CN)_4^-$, $PtCl_4^{2-}$ and $PdCl_4^{2-}$] or neutral [*e.g.* $Pt(NH_3)_2Cl_2$] complexes. These may accept an additional ligand to give square pyramidal coordination or two ligands to give octahedral coordination. The additional ligands are normally more weakly bound. Pt(IV) has a d^6 configuration and forms stable octahedral complexes, such as $PtCl_6^{2-}$, with six equivalent covalently bound ligands.

The kinetic and thermodynamic stability of these complexes depends on the protein ligands, buffer, pH and salting in/out agent (Petsko *et al.*, 1978). Anionic groups do not readily react with anionic reagents, such as RS^-, but are attacked more readily by neutral nucleophiles such as RSH, R-imidazole or RNH_2. The inert cationic group $Pt(NH_3)_4^{2+}$ is most likely to form electrostatic complexes with anionic groups, such as carboxylate. The neutral $Pt(NH_3)_2Cl_2$ molecule, however, can penetrate into hydrophobic areas but requires a stronger nucleophile such as RS^-. In acidic and neutral solutions, $PtCl_4^{2-}$ reacts most commonly with methionine (Figs. 12.1.5.4 and 12.1.5.5), cystine (disulfide) (Fig. 12.1.5.6), N-termini and histidine to form stable complexes. However, methionine reacts faster than histidine. Thus, it is possible to use time as a variable to define specificity. The most popular platinum reagents are listed in Table 12.1.5.4.

In aqueous solution, the square-planar complex $AuCl_4^-$ is hydrolysed to $Au(OH)_4^-$ in about one hour, or in the presence of

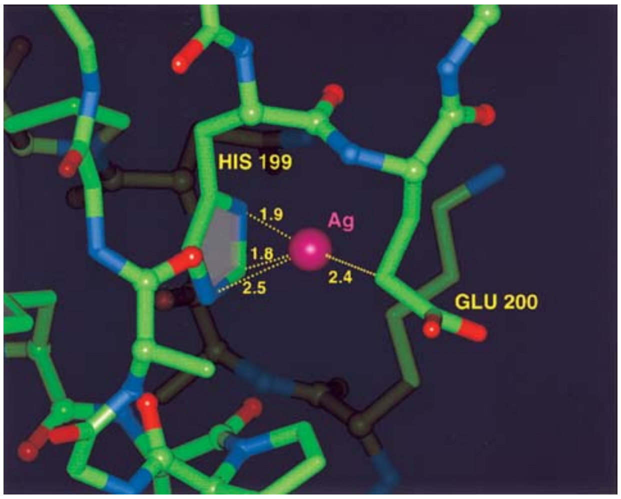


Figure 12.1.5.3
The binding of a silver ion to immunoglobulin Fab (2FB4). The positions of the ligands in the parent crystals are shown, and these must move in the complex to coordinate the silver ion.

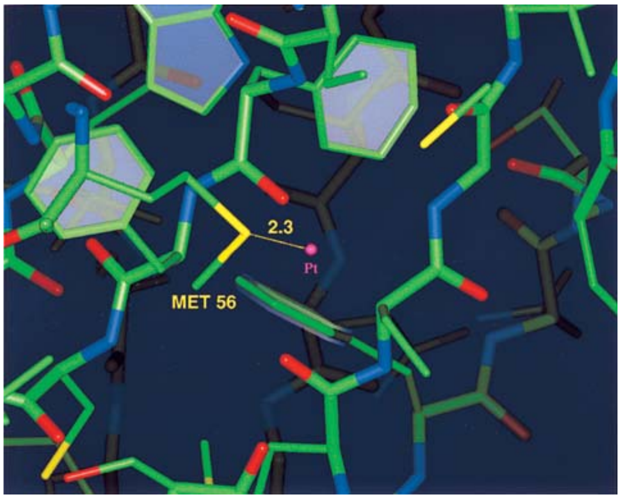


Figure 12.1.5.4
The binding of $PtCl_4^{2-}$ through a methionine in azurin (1AZU).

a protein, reduced to Au(I) by methionine. In ammonium sulfate it probably exists as $AuCl_3(NH_3)$, $AuCl_2(NH_3)_2^+$ and $Au(NH_3)_4^{3+}$. In contrast, $Au(CN)_2^-$ is more stable and normally binds electrostatically. However, on occasions at pH > 6.0, the $Au(CN)_2^-$ complex has bound to cysteine residues by nucleophilic displacement reactions.

Osmium resembles platinum in many ways and typically acts as a class B metal. It occurs in all oxidation states from 0 to VIII, but most usually in III, as in K_3OsCl_6; in IV, as in K_2OsCl_6; in VI, as in $K_2OsO_2(OH)_4$; and in VIII, as in osmium tetraoxide, OsO_4. Higher-oxidation-state compounds tend to be reduced to $OsO_2(OH)_2$ in most crystallization solutions and in the presence of ammonia or halide ion they can become further reduced to cationic or anionic complexes, such as $Os(NH_3)_6^{3+}$ or $OsCl_6^{2-}$. Anionic complexes may be substituted by histidine residues at pH > 7.0 or bound as ion pairs by histidine at pH < 7.0 or protonated amino groups. Cationic complexes tend to bind to negatively charged residues *via* electrostatic interactions.

Iridium is found in all oxidation states from II to VI but commonly exists in III, as in K_3IrCl_6, and IV, as in $(NH_4)_2IrCl_6$. Ir(III) is similar to *rhodium*(III) and is found in a variety of cationic, uncharged and anionic complexes. All Ir(III) complexes are kinetically inert, whereas most anionic complexes of Rh(III) are labile. Ir(IV) is commonly found as the hexahalo complexes IrX_6^{2-} (except iodine), which are also fairly kinetically inert. Cationic [*e.g.* $Ir(NH_3)_6^{3+}$], neutral (*i.e.* $IrCl_3$) and anionic (*i.e.* $IrCl_6^{2-}$) species have proved useful in forming derivatives of protein crystals.

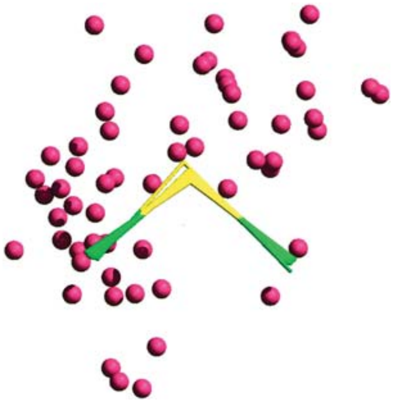


Figure 12.1.5.5
The relative positions of methionine side chains (carbon: green; sulfur: yellow) in the parent crystals to the binding of platinum (pink) of $PtCl_4^{2-}$. The methionine side chains have been least-squares fitted.

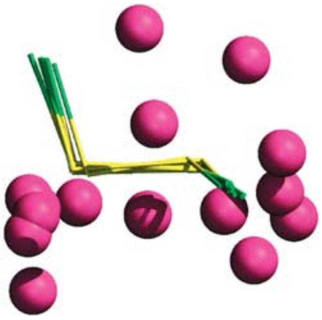


Figure 12.1.5.6
The relative positions of cystine disulfide bridges (carbon: green; sulfur: yellow) in the parent crystals to the binding of platinum (pink) of $PtCl_4^{2-}$. The cystine side chains have been least-squares fitted, and only those with torsion angles in the range 99.7 ± 8.3° have been used.

12.1.5.4. Electrostatic binding of heavy-atom anions

Positively charged groups of proteins, such as the α-amino terminus, ε-amino of lysine, guanidinium of arginine and imadazolium of histidine, may form ion pairs with heavy-atom anionic complexes. For example, HgI_4^{2-} and HgI_3^- can bind through electrostatic interactions. Anionic metal cyanide complexes tend to be more resistant to substitution and consequently interact electrostatically on most occasions. For example, $Pt(CN)_4^{2-}$ binds at several sites involving lysine or arginine residues in proteins (Fig. 12.1.5.7). $Pt(CN)_4^{2-}$ and $Au(CN)_2^-$ can act as inhibitors by binding at coenzyme phosphate sites.

12.1.5.5. Hydrophobic heavy-atom reagents

Since many heavy-atom reagents are hydrophilic, most interactions occur at the protein surface. However, substitution, addition or removal of the non-heavy-atom component(s) of the reagent can alter the hydrophilic–hydrophobic balance and lead to penetration of the core. For example, anionic complexes such as $HgCl_4^{2-}$ and $PbCl_6^{2-}$ are hydrophilic and would not normally enter the protein core, although organometallics, such as *R*HgCl

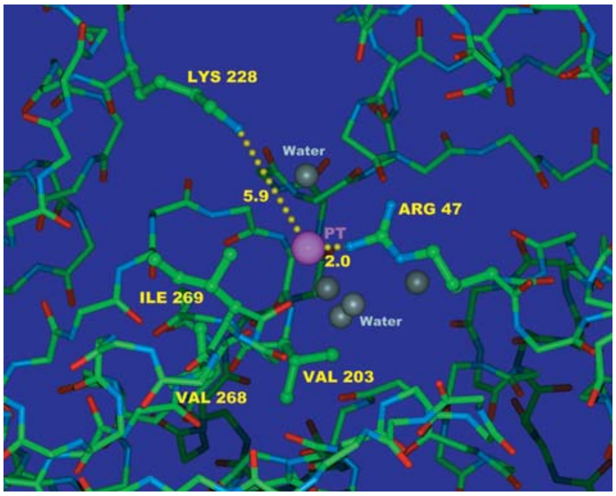


Figure 12.1.5.7
The binding of $Pt(CN)_4^{2-}$ to aldose dehydrogenase (8ADH).

and R_3PbCl (R = aliphatic or aromatic), are much more hydrophobic and can do so.

Hydrophobic organomercury compounds of the general formula $R$Hg$X$, where R is an aliphatic or aromatic organic group, react with sulfhydryls through displacement of X. When X is PO_4^{3-}, SO_4^{2-} or NO_3^-, the bond is ionic, making the formation of the cation $R\mathrm{Hg}^+$ easier. R is often chosen to be a small aliphatic group (*e.g.* CH_3, C_2H_5). However, the presence of a benzene ring enhances the stability of the heavy-atom reagent. Careful selection of the X group can assist penetration into the hydrophobic core. The hydrophobicity of X follows the order

$$PO_4^{3-} < NO_3^- < Cl < Br < I < R.$$

$R$Hg$R$ (R = aliphatic or aromatic) compounds also bind sulfhydryl residues in hydrophobic regions. The mechanism of reaction of methylphenylmercury with buried sulfhydryl groups may involve fast dissolution in the hydrophobic interior of the protein followed by a slow reaction with neighbouring sulfhydryl residues (Abraham *et al.*, 1983). They are difficult to prepare in aqueous solutions; an aprotic solvent, such as acetonitrile, can improve solubility, but this is not normally a problem in high concentrations of organic components, such as PEG, MPD or ethanol.

Inert gases were first used in the analysis of myoglobin. Schoenborn *et al.* (1965) discovered that the hydrophobic site that bound HgI_3^- also bound a xenon atom at 2.5 atmospheres. They proposed that this may be a general way of producing heavy-atom derivatives of proteins. Recently, there has been increasing interest in this idea, which has now been developed to produce well defined derivatives of a wide range of different proteins. Crystals are subjected to high gas pressures. Xenon requires about 10 atmospheres in order to get saturated binding sites. Krypton binds much less strongly and requires around 60 atmospheres. Since the binding of both inert gases is reversible, it is necessary to keep the protein crystals in a gaseous environment in a specialized pressure cell. Such pressure cells have been developed by Schiltz (1997) at LURE. Xenon binds to hydrophobic cavities, with little conformational change and a retention of isomorphism in crystals. Krypton binds at the same sites as xenon, but since it is lighter and needs higher pressure it has been exploited less by protein crystallographers. However, it has a well defined K edge at around 1 Å and so has attractions for multiple-wavelength anomalous dispersion.

12.1.5.6. Iodine

In addition to their use in isomorphous replacement, iodine derivatives of crystalline proteins have been prepared as tyrosine or histidine markers to assist main-chain tracing and to act as a probe for surface residues. The order of reactivity towards these reactive residues is

$$\mathrm{Tyr} > \mathrm{His} \gg \mathrm{Trp}.$$

I_3^-, I^-, I^+ and I_2 can be generated by several different methods. An equimolar solution of KI/I_2 or NaI/I_2 in 5% (v/v) ethanol/water solution is often used to generate the anionic species I_3^- and I^-. An oxidizing agent, such as chloramine T, can be added to KI, typically in a concentration ratio of 1:50; alternatively, polystyrene beads derivatized with *N*-chlorobenzene sulfonamide can be used with NaI. Similarly, the addition of excess KI to ICl or OI^- will generate I_3^-, I^- and I^+. To avoid oxidation of iodine solutions, the pH should be less than 5.0. To avoid cracking the crystals, it may be necessary to increase the iodine concentration very slowly and to wash the derivatized crystals in the mother liquor in order to remove free I_2. Mono- or di-iodination of tyrosines can cause disruption of the protein structure either because of the larger size or the breaking of hydrogen bonds due to lowering of the pK_a of the phenolic hydroxyl.

12.1.5.7. Polynuclear reagents

The structure determination of large multicomponent systems such as the 50S ribosomal subunit (Yonath *et al.*, 1986) or the nucleosome core particle (O'Halloran *et al.*, 1987) requires the addition of reagents with a greater number of electrons, preferably in a compact polynuclear structure. Such reagents may be either cluster compounds or multimetal centres having metal–metal bonds.

Polynuclear reagents should preferably be covalently bound to one or a few specific sites, either first in solution or later in the crystals. Spacers of differing length can be inserted into the reagent to increase accessibility. Their low solubility in aqueous solutions can often be overcome by dissolving them in an apolar solvent (*e.g.* acetonitrile). Tetrakis(acetoxymercurio)methane (TAMM) and di-*m*-iodobis(ethylenediamine)diplatinum(II) nitrate (PIP) have better solubility in aqueous solutions than other polynuclear heavy-atom compounds.

Polynuclear heavy-atom reagents give an enhanced signal-to-noise ratio in low-resolution MIR studies, but this advantage is offset by the fall-off in scattering amplitude that arises from interference of diffracted waves at higher resolution. In the nucleosome core particle, the scattering reached 50% of its zero-angle value at 7.0 Å, while the relative drop for a single heavy atom was 10% (O'Halloran *et al.*, 1987). Cluster and multimetal reagents that have been successfully employed in protein structure determinations have been reviewed by Thygesen *et al.* (1996).

12.1.6. Metal-ion replacement in metalloproteins

The metal-ion cofactor can sometimes be displaced by dialysis or diffusion by a heavy-atom solution, but usually the cofactor is first removed by a chelating agent (*e.g.* EDTA) or by acidification. These are best carried out on the crystals. Alternatively, the metal can be substituted by biosynthesis of the metalloprotein under enriched conditions of the substituting metal, an approach which has been successful in displacing zinc with cobalt and other lighter metals.

The metal ions are best substituted by a metal of similar character and radius. Thus, calcium is an A-group metal which prefers ligands containing oxygen atoms that may originate from carboxylic, carboxyamide, hydroxyl, main-chain carbonyl groups and water molecules. Divalent alkaline earth metal ions (*e.g.* Sr^{2+}, Ba^{2+}) or trivalent lanthanide ions can bind at calcium sites but can give very different coordination geometry and stability. Nd^{3+} and Sm^{3+} can displace some Ca^{2+} ions with negligible change in structure (Fig. 12.1.6.1). On the other hand, zinc has a relatively small ionic radius and is more polarizing. Structural zinc atoms are often tetrahedrally coordinated by cysteine residues, while those at active sites frequently bind histidine, often in association with a water molecule and/or carboxylate ligands. Cadmium or mercury can replace zinc, but often with a conformational change leading to lack of isomorphism.

12.1.7. Analogues of amino acids

Attempts to replace amino acids by heavy-atom substituted synthetic analogues with a similar charge and shape have not

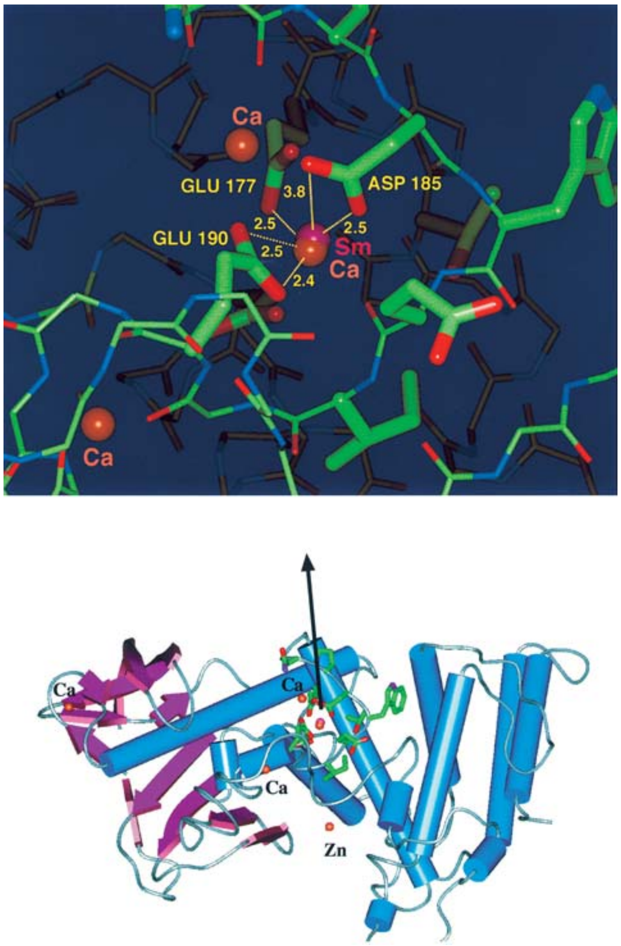


Figure 12.1.6.1
The displacement of calcium by samarium in thermolysin. The samarium of the heavy-atom derivative is shown superposed on the parent crystal structure.

proved successful in large proteins, although a selenocystine was used successfully in the analysis of oxytocin (Wood *et al.*, 1986). However, the production of proteins labelled by selenium using biological substitution of selenomethionine (SeMet) for methionine (Hendrickson, 1985) has been stimulated by multiple-wavelength anomalous dispersion (MAD) (Hendrickson *et al.*, 1990). Methionine biosynthesis is blocked in the cells in which the protein is produced and SeMet is substituted for Met in the growth medium. The generality of the labelling scheme for proteins is the root of its success, as discussed by Doublié (1997).

SeMet has been incorporated into proteins expressed in *Escherichia coli* strains that are auxotrophic for Met [strain DL421 (Hendrickson *et al.*, 1990); strain B834 (Leahy *et al.*, 1994); strain LE392 (Ceska *et al.*, 1996)]. Nearly complete incorporation has also been reported in non-auxotrophic bacterial strains, *E. coli* strain XA90 (Labahn *et al.*, 1996), in a mammalian cell line (Lustbader *et al.*, 1995) and in baculovirus-infected insect cells (Chen & Bahl, 1991). Usually, somewhat higher than normal concentrations of disulfide reducing agents, such as dithiothreitol or mercaptoethanol, are sufficient to protect SeMet from air oxidation to the selenoxide, although crystallization in an inert atmosphere may be necessary. Proteins usually have SeMet substituted for Met at levels approaching 100%. The cells are viable and the proteins are functional.

Site-directed mutagenesis offers an alternative approach for the introduction of specific heavy-atom binding sites. A common procedure is to replace residue(s) in the variable part of the primary structure with cysteine. The selection of the residue to mutate in a protein of unknown structure remains a challenge.

Although selenocysteine is toxic to cells, cysteine auxotrophic strains, in which proteins can be synthesized with the seleno derivative, have been developed (Miller, 1972; Muller *et al.*, 1994). The bacteria are grown under limiting amounts of cysteine with no other sulfur source. They are induced for 10 min and then resuspended in selenocysteine for a 3 h incubation. The protein is purified with a reducing agent. In general, the substitution at the selenocysteine seems to be less satisfactory than selenomethionine, with occupancy often as low as 20%.

Budisa *et al.* (1997) have experimented on incorporating a range of novel amino-acid analogues using *in vitro* suppression. This is achieved by suppressing the stop colons and engineering tRNA synthases to incorporate the analogue. Possible candidates are telluromethionine, 5-bromotryptophan, 5-iodotryptophan, selenotryptophan and tellurotryptophan. The bioincorporation of TeMet into derivatized crystals did not greatly affect their stability in buffer solutions and to X-radiation. Isomorphism was maintained despite the C—Te bond being longer than C—Se or C—S. TeMet crystals are not as suitable for MAD analysis as SeMet crystals due to the 0.3 Å absorption edge of tellurium. The method is restricted to methionine residues located in the hydrophobic regions, since solvent accessibility may cause undefined chemical reactions with the highly reactive C—Te side chain. Thus the protein must be expressed in the folded form.

12.1.8. Use of the heavy-atom data bank to select derivatives

The heavy-atom data bank is probably best exploited by first investigating the most commonly used heavy-atom reagents with a view to obtaining mercury, platinum and uranyl derivatives that tend to bind at different sites. The most common reagents (Table 12.1.5.1) can first be selected and tested for suitability in terms of amino-acid sequence, pH, buffer and salt. If there are many sulfhydryls, several mercurials might be exploited, or if there are several methionines, other platinum agents might be investigated. A high pH would argue against use of uranyl due to the insolubility of hydroxides; the presence of ammonium sulfate would argue for as low a pH as possible. The presence of citrate would imply changing the buffer for acetate if A-group metals, such as uranium or lanthanides, were to be used.

For each heavy-atom agent, the conditions of its previous use can be checked against the conditions of crystallization in the current study. Conversely, the database can be interrogated for reagents that have been used in similar conditions. In each case, derivatives that maximize the variety of ligands should be exploited.

The time of soak should be first set according to previous experience indicated in the database. However, the progress of heavy-atom substitution needs to be monitored by checking for change of colour, transparency or cracking. If cracking and disruption of the crystals occurs quickly, a less reactive reagent can be tried, and, conversely if substitution is insufficient, a more reactive reagent can be tried. If there are several cysteines, different derivatives can be obtained with mercurials of different size and hydrophobicity. In each circumstance, the data bank should provide useful information to assist in choosing reagents.

Please keep information about the heavy-atom binding sites and the heavy-atom structure-factor amplitudes. They should be submitted to the Protein Data Bank.

We thank all those who have generously sought out and sent us details of the heavy-atom binding sites in their derivatives, and the ICRF and the Wellcome Trust for support.

References

Abraham, D. J., Phillips, S. E. V. & Kennedy, P. E. (1983). *Methylphenylmercury: a novel heavy atom reagent for protein crystallography. J. Mol. Biol.* **170**, 249–252.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 534–552.

Blake, C. C. F. (1968). *The preparation of isomorphous derivatives.* In *Advances in Protein Chemistry,* Vol. 23, pp. 59–120. New York and London: Academic Press.

Blundell, T. L. (1968). Unpublished results.

Blundell, T. L. & Jenkins, J. A. (1977). *The binding of heavy metals to proteins. Chem. Soc. Rev. (London)*, **6**, 139–171.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography.* New York: Academic Press.

Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neuefeind, T., Moroder, L. & Huber, R. (1997). *Bioincorporation of telluromethionine into proteins: a promising new approach for X-ray structure analysis of proteins. J. Mol. Biol.* **271**, 1–8.

Carvin, D. G. A. (1986). Unpublished results.

Carvin, D. G. A., Islam, S. A., Sternberg, M. J. E. & Blundell, T. L. (1991). *Isomorphous Replacement and Anomalous Scattering*. Warrington: Daresbury Laboratory.

Ceska, T. A., Sayers, J. R., Stier, G. & Suck, D. (1996). *A helical arch allowing single-stranded DNA to thread through t5 5′-exonuclease. Nature (London)*, **382**, 90–93.

Chen, W. & Bahl, O. P. (1991). *Recombinant carbohydrate and selenomethionyl variants of human choriogonadotropin. J. Biol. Chem.* **266**, 8192–8197.

Doublié, S. (1997). *Preparation of selenomethionyl proteins for phase determination. Methods Enzymol.* **276**, 523–530.

Ely, K. R., Girling, R. L., Schiffer, M., Cunningham, D. E. & Edmundson, A. B. (1973). *Preparation and properties of crystals of a Bence–Jones dimer with mercury inserted into the interchain disulphide bond. Biochemistry*, **12**, 4233–4237.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. E. (1994). *Biological Macromolecular Crystallization Database, version 3.0: new features, data and the NASA archive for protein crystal growth data. Acta Cryst.* D**50**, 408–413.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The method of isomorphous replacement for protein crystallography. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Hendrickson, W. A. (1985). *Analysis of protein structure from diffraction measurement at multiple wavelengths. Trans. Am. Crystallogr. Assoc.* **21**, 11–21.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J.* **9**, 1665–1672.

Islam, S. A., Carvin, D., Sternberg, M. J. E. & Blundell, T. L. (1998). *HAD, a data bank of heavy-atom binding sites in protein crystals: a resource for use in multiple isomorphous replacement and anomalous scattering. Acta Cryst.* D**54**, 1199–1206.

Labahn, J., Scharer, O. D., Long, A., Ezaz-Nikpay, K., Verdine, O. L. & Ellenberger, T. E. (1996). *Structural basis for the excision repair of alkylation-damaged DNA. Cell*, **86**, 321–329.

Leahy, D. J., Erickson, H. P., Aukhil, I., Joshi, P. & Hendrickson, W. A. (1994). *Crystallization of a fragment of human fibronectin: introduction of methionine by site-directed mutagenesis to allow phasing via selenomethionine. Proteins Struct. Funct. Genet.* **19**, 48–54.

Lipka, J. J., Lippard, S. J. & Wall, J. S. (1976). *Visualisation of polymercurimethane-labelled fd bacteriophage in the scanning transmission electron microscope. Science*, **206**, 1419–1421.

Lipscomb, W. N., Reeke, G. N., Hartsuck, J. A., Quiocho, F. A. & Bethge, P. H. (1970). *The structure of carboxypeptidase a. VIII. Atomic interpretation at 0.2 nm resolution, a new study of the complex of glycyl-L-tyrosine with CPA, and mechanistic deductions. Philos. Trans. R. Soc. London Ser. B*, **257**, 177–214.

Lustbader, J. W., Wu, H., Birken, S., Pollak, S., Kolks-Gawinowicz, M. A., Pound, A. M., Austen, D., Hendrickson, W. A. & Canfield, R. E. (1995). *The expression, characterization and crystallization of wild-type and selenomethionyl human chorionic gonadotrophin. Endocrinology*, **136**, 640–650.

Miller, J. H. (1972). *Cysteine auxotrophic strains, in which proteins can be synthesised with the seleno derivative.* In *Experiments in Molecular Genetics.* Cold Spring Harbour Laboratory Press.

Muller, S., Senn, H., Gsell, B., Vetter, W., Baron, C. & Bock, A. (1994). *The formation of diselenide bridges in proteins by incorporation of selenocysteine residues: biosynthesis and characterization of (Se)2-thioredoxin. Biochemistry*, **33**, 3404–3412.

O'Halloran, T. V., Lippard, S. J., Richmond, T. J. & Klug, A. (1987). *Multiple heavy-atom reagents for macromolecular X-ray structure determination application to the nucleosome core particle. J. Mol. Biol.* **194**, 705–712.

Petsko, G. A., Phillips, D. C., Williams, R. J. P. & Wilson, I. A. (1978). *On the protein crystal chemistry of chloroplatinite ions: general principles and interactions with triose phosphate isomerase. J. Mol. Biol.* **120**, 345–359.

Ringe, D., Petsko, G. A., Yamakura, F., Suzuki, K. & Ohmori, D. (1983). *Structure of iron superoxide dismutase from Pseudomonas ovalis at 2.9 Å resolution. Proc. Natl Acad. Sci. USA*, **80**, 3879–3883.

Sasaki, K., Dockerill, S., Adamiak, D. A., Tickle, I. J. & Blundell, T. L. (1975). *X-ray analysis of glucagon and its relationship to receptor binding. Nature (London)*, **257**, 751–757.

Schiltz, M. (1997). *Xenon & krypton at LURE.* http://www.lure.u-psud.fr/sections/Xenon/XENON_ENG.HTM.

Schoenborn, B. P., Watson, H. C. & Kendrew, J. C. (1965). *Binding of xenon to sperm whale myoglobin. Nature (London)*, **207**, 28–30.

Sigler, P. B. & Blow, D. M. (1965). *A means of promoting heavy atom binding in protein crystals. J. Mol. Biol.* **14**, 640–644.

Sigler, P. B., Blow, D. M., Matthews, B. W. & Henderson, R. (1968). *Structure of crystalline alpha-chymotrypsin II. A preliminary report including a hypothesis for the activation mechanism. J. Mol. Biol.* **35**, 143–164.

Sperling, R., Burstein, Y. & Steinberg, I. Z. (1969). *Selective reduction and mercuration of cysteine IV–V in bovine pancreatic ribonuclease. Biochemistry*, **8**, 3810–3820.

Sperling, R. & Steinberg, I. Z. (1974). *Simultaneous reduction and mercuration of disulphide bond A6–A11 of insulin by monovalent mercury. Biochemistry*, **13**, 2007–2013.

Thygesen, J., Weinstein, S., Franceschi, F. & Yonath, A. (1996). *The suitability of metal clusters for phasing in macromolecular crystallography of large macromolecular assemblies. Structure*, **4**, 513–518.

Tsernoglou, D. & Petsko, G.-A. (1976). *The crystal structure of a post-synaptic neurotoxin from sea snake at 2.2 Å resolution. FEBS Lett.* **68**, 1–4.

Wood, S. P., Tickle, I. J., Treharne, A. M., Pitts, J. E., Mascarenhas, Y., Li, J. Y., Husain, J., Cooper, S., Blundell, T. L., Hruby, V. J., Buku, A., Fischman, A. J. & Wyssbrod, H. R. (1986). *Crystal structure analysis of deamino-oxytocin: conformational flexibility and receptor binding. Science*, **232**, 633–636.

Yonath, A., Saper, M. A., Makowski, I., Mussig, J., Piefke, J., Bartunik, H. D., Bartels, K. S. & Wittmann, H. G. (1986). *Characterization of single crystals of the large ribosomal particles from bacillus stearothermophilus. J. Mol. Biol.* **187**, 633–636.

International Tables for Crystallography (2012). Vol. F, Chapter 12.2, pp. 327–332.

Chapter 12.2. Locating heavy-atom sites

M. T. Stubbs and R. Huber

12.2.1. The origin of the phase problem

Once a native data set has been collected, the next task is the solution of the structure. There is one major hurdle: the phase problem. To study objects at the atomic level, we must utilize waves with a wavelength in the ångström range, *i.e.* X-radiation. X-rays interact with electrons and so provide an image of the electron distribution of the sample. Unfortunately, X-rays are refracted by matter only very weakly, and so it is not possible to construct a lens to view molecules at atomic dimensions.[1]

As shown in Chapter 2.1, the diffraction $F(\mathbf{S})$ obtained from an electron-density distribution $\rho(\mathbf{r})$ is given by

$$F(\mathbf{S}) = \textstyle\int \rho(\mathbf{r}) \exp\{2\pi i \mathbf{r} \cdot \mathbf{S}\}\, \mathrm{d}^3\mathbf{r},$$

where $\mathbf{S}$ is perpendicular to the scattered wave and $|\mathbf{S}| = 2\sin\theta/\lambda$; θ is the scattering angle and λ is the wavelength. The diffraction pattern is a *Fourier transform* of the electron density. If we have a crystal with cell parameters a, b and c, then the Laue diffraction conditions require that $\mathbf{S}$ lies on a reciprocal lattice such that $\mathbf{S} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$, where $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$ are the reciprocal-lattice vectors, and h, k and l are the integer indices of the diffracted beam.

$$F(hkl) = V \sum_{xyz} \rho(xyz) \exp\{2\pi i(hx + ky + lz)\},$$

where V represents the volume of the unit cell, and x, y and z are the fractional coordinates within that cell in the directions of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$.

Since the diffraction pattern is a Fourier transform of the electron density, it follows that the electron density is an inverse Fourier transform of the diffraction pattern:

$$\rho(\mathbf{r}) = \textstyle\int F(\mathbf{S}) \exp\{-2\pi i \mathbf{r} \cdot \mathbf{S}\}\, \mathrm{d}^3\mathbf{S},$$

$$\rho(xyz) = (1/V) \sum_{hkl} F(hkl) \exp\{-2\pi i(hx + ky + lz)\}.$$

Thus it should be mathematically straightforward to calculate the electron density from the diffraction pattern. This is, unfortunately, not the case. The function $F(\mathbf{S})$ describing the diffracted rays is a complex function with a *magnitude* $|F(\mathbf{S})|$ and a *phase* $\varphi(\mathbf{S})$. The diffraction experiment measures the *intensities* $I(\mathbf{S})$, however; the relationship between $I(\mathbf{S})$ and $F(\mathbf{S})$ is:

$$I(\mathbf{S}) = F^*(\mathbf{S}) \cdot F(\mathbf{S}) = |F(\mathbf{S})|^2,$$

where $F^*(\mathbf{S})$ is the complex conjugate of $F(\mathbf{S})$. The measured intensities are related directly to the magnitudes of the diffracted beams; the phase information, however, is lost (Fig. 12.2.1.1): this is the origin of the *phase problem*.

There are essentially four ways of overcoming the phase problem (Fig. 12.2.1.2):

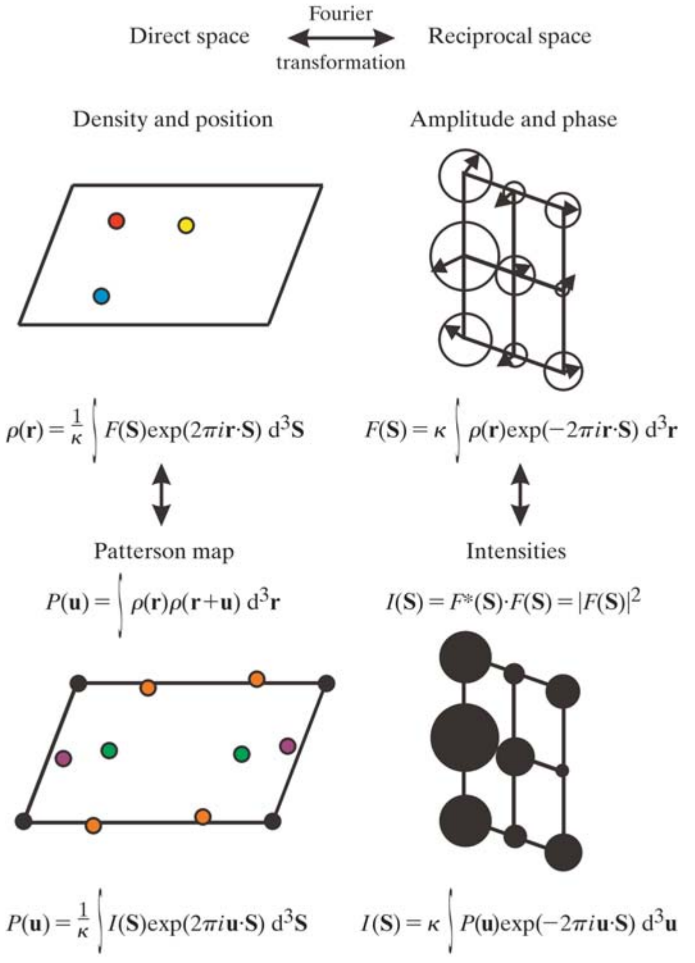
(1) the use of isomorphous replacement to influence the diffraction pattern, thereby revealing information about the phases (this chapter);

(2) Patterson search techniques, which in essence allow modelling of the phase distribution of an unknown crystal based upon that of a known molecule (Part 13);

(3) the use of anomalous dispersion, which shifts the imaginary component of $F(\mathbf{S})$, allowing an experimental measurement of the phase (Part 14); and

(4) direct methods, which make use of probabilistic relationships between different diffracted rays (Part 16).

The method of isomorphous replacement, by which the first macromolecular structures were solved (Green *et al.*, 1954), remains the most widely used technique for *ab initio* structure



Direct space
Fourier transformation
Reciprocal space
Density and position
Amplitude and phase
$\rho(\mathbf{r}) = \frac{1}{\kappa}\int F(\mathbf{S})\exp(2\pi i\mathbf{r}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{S}$
$F(\mathbf{S}) = \kappa\int \rho(\mathbf{r})\exp(-2\pi i\mathbf{r}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{r}$
Patterson map
$P(\mathbf{u}) = \int \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\,\mathrm{d}^3\mathbf{r}$
Intensities
$I(\mathbf{S}) = F^*(\mathbf{S})\cdot F(\mathbf{S}) = |F(\mathbf{S})|^2$
$P(\mathbf{u}) = \frac{1}{\kappa}\int I(\mathbf{S})\exp(2\pi i\mathbf{u}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{S}$
$I(\mathbf{S}) = \kappa\int P(\mathbf{u})\exp(-2\pi i\mathbf{u}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{u}$

Figure 12.2.1.1
Relationships in diffraction space. The diffraction pattern of an object is the Fourier transform of the electron density, consisting of both amplitudes and phases. What we measure in an X-ray diffraction experiment, however, are the diffracted intensities; the phase information is lost. The Fourier transform of the intensities results in the Patterson map, which is related to the electron density as follows. For any two atoms in the structure, the vector between them, centred at the origin, has a value corresponding to the product of their densities. Thus, the red atom and the yellow atom result in the orange cross vectors, the red and blue atoms result in the magenta cross vectors, and the yellow and blue atoms result in green cross vectors. As each atom has a 'cross vector' to itself, a large peak is found at the origin; the Patterson map is centrosymmetric.

---

[1] This is not strictly true, due to advances in electron microscopy (see Part 19); however, this method has yet to find the universality enjoyed by X-ray crystallography.

Copyright © 2012 International Union of Crystallography

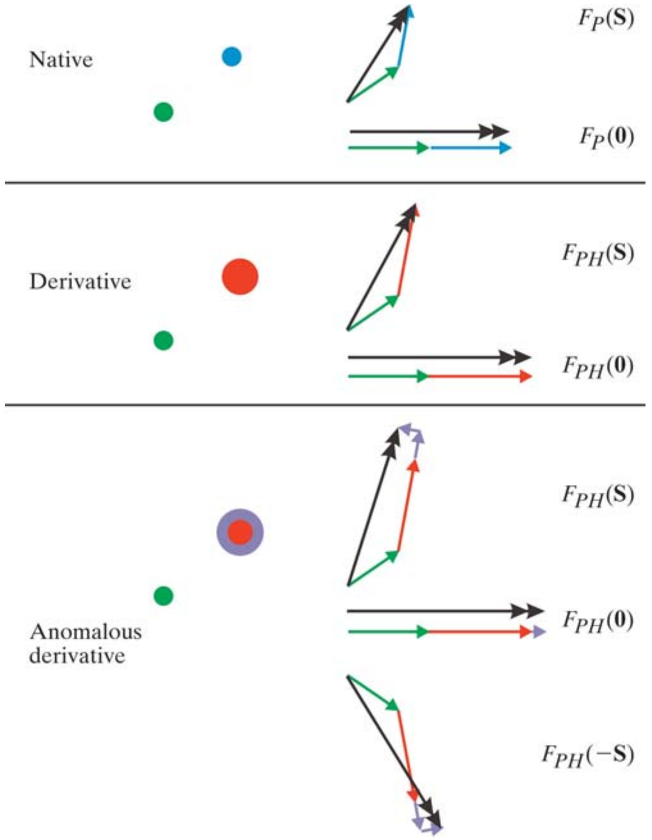


Figure 12.2.1.2
The effect of introducing a heavy atom or anomalous scatterer. The native two-atom structure gives rise to two diffraction vectors (green and blue) of equal magnitude but different phase (see Chapter 2.1), with a resultant diffraction vector F_P (black). Isomorphous replacement of the blue atom by the larger red one gives rise to a diffraction vector of greater magnitude but equivalent phase (red), causing a change in the resultant magnitude F_{PH} (and hence the intensity) *and* in the phase. Introduction of an anomalous scatterer results in a phase shift (lilac) of the diffraction vector, resulting in differing amplitudes and phases for $F_{PH}(\mathbf{S})$ and $F_{PH}(-\mathbf{S})$.

determination, although the availability of synchrotrons, with their facility for selecting a desired wavelength, and molecular-biology techniques that allow the direct introduction of anomalous scatterers, such as selenium or tellurium, into the protein of interest (Hendrickson *et al.*, 1990; Budisa *et al.*, 1997) have proven that multiple anomalous dispersion is an exceptionally powerful technique for the solution of novel structures. Patterson search techniques (Rossmann, 1972) are ideal if a similar macromolecular structure is already known, while direct methods are more-or-less confined to very high resolution data (Sheldrick, 1990).

In order to obtain phase information from isomorphous replacement (or from anomalous dispersion), it is necessary to locate the atomic positions of the heavy-atom (or anomalous) scatterers.

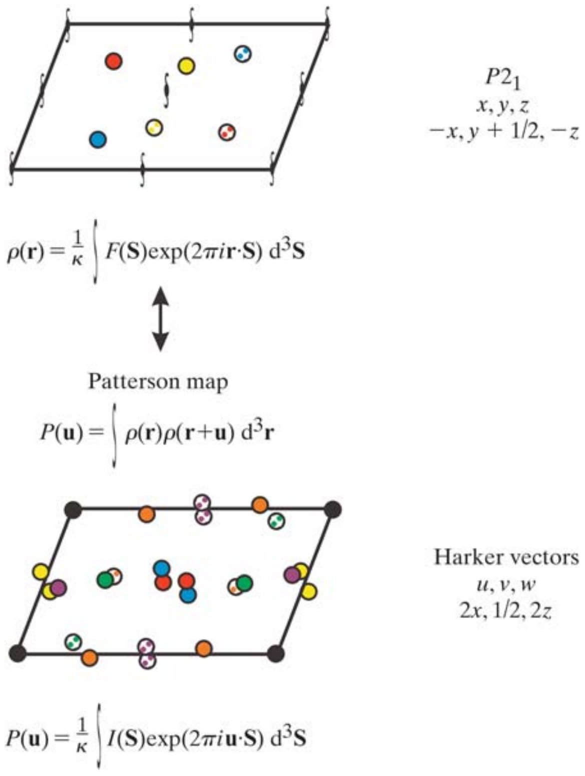

$P2_1$
x, y, z
$-x, y+1/2, -z$
$\rho(\mathbf{r}) = \frac{1}{\kappa}\int F(\mathbf{S})\exp(2\pi i\mathbf{r}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{S}$
Patterson map
$P(\mathbf{u}) = \int \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\,\mathrm{d}^3\mathbf{r}$
Harker vectors
u, v, w
$2x, 1/2, 2z$
$P(\mathbf{u}) = \frac{1}{\kappa}\int I(\mathbf{S})\exp(2\pi i\mathbf{u}\cdot\mathbf{S})\,\mathrm{d}^3\mathbf{S}$

Figure 12.2.2.1
The Patterson map with symmetry. When the crystal unit cell contains more than one molecule, then additional cross vectors will be formed between differing molecules. If these are related by crystallographic symmetry, there is a geometrical relationship between cross peaks. In this diagram, the peaks of Fig. 12.2.1.1 are supplemented by those between atoms of symmetry-related molecules. The red, yellow and blue peaks of the resulting Patterson function represent those between same atoms (*i.e.* red to red, yellow to yellow and blue to blue) related by symmetry. These peaks are found on a Harker section.

12.2.2. The Patterson function

Although the set of measured intensities contains no information regarding the phases, the Fourier transform of the intensities, the so-called Patterson function, contains valuable information. Patterson (1934) showed that the inverse Fourier transform of the intensity,

$$P(uvw) = (1/V)\sum_{hkl} I(hkl)\exp\{-2\pi i(hu+kv+lw)\},$$

is related to the electron density by

$$P(\mathbf{u}) = \textstyle\int \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\ \mathrm{d}^3\mathbf{r}.$$

The Patterson function $P(\mathbf{u})$ is an autocorrelation function of the density. For every vector $\mathbf{u}$ that corresponds to an interatomic vector, $P(\mathbf{u})$ will contain a peak (Fig. 12.2.1.1). These are some properties of the Patterson function:

(1) Every atom makes an 'interatomic vector' with itself, and therefore the *origin peak*, $P(\mathbf{0}) = \sum \rho^2(\mathbf{r})$, dominates the Patterson function. This origin peak can be 'removed' through subtraction of the average intensity from $I(hkl)$ before Fourier transformation.
(2) For every vector between $\rho_i(\mathbf{r}_i)$ and $\rho_j(\mathbf{r}_j)$, the same value (*i.e.* their product) is found for $\rho_j(\mathbf{r}_j)$ to $\rho_i(\mathbf{r}_i)$, and so the Patterson map is *centrosymmetric*.
(3) For a structure consisting of n atoms, there are $n(n-1)/2$ cross vectors, and so the Patterson function is extremely crowded.

For simple crystals, the Patterson map can be used to solve the structure directly. For macromolecular structures, the Patterson map provides a vehicle for solving the phase problem.

If the crystal contains rotational symmetry elements, then the cross vectors between $\rho_i(\mathbf{r}_i)$ and its symmetry mate lie on a plane perpendicular to the symmetry axis – the *Harker section* (Harker, 1956). By way of example, the space group $P2_1$ has two symmetry-related positions (Fig. 12.2.2.1),

$$(x, y, z) \text{ and } (-x, y + \tfrac{1}{2}, -z).$$

Cross vectors between symmetry-related points will therefore have the form

$$(2x, \tfrac{1}{2}, 2z),$$

i.e. all cross vectors lie on the plane $v = \frac{1}{2}$. For space group $P2_12_12_1$, the general coordinates

$$(x, y, z), (x + \tfrac{1}{2}, -y + \tfrac{1}{2}, -z), (-x + \tfrac{1}{2}, -y, z + \tfrac{1}{2}),$$
$$(-x, y + \tfrac{1}{2}, -z + \tfrac{1}{2})$$

give rise to cross vectors

$$(\tfrac{1}{2}, 2y + \tfrac{1}{2}, 2z), (2x + \tfrac{1}{2}, 2y, \tfrac{1}{2}), (2x, \tfrac{1}{2}, 2z + \tfrac{1}{2}),$$

i.e. there are three Harker sections: $u = \frac{1}{2}$, $v = \frac{1}{2}$ and $w = \frac{1}{2}$. Peaks occurring on the Harker sections must reduce to a self-consistent set of coordinates (x, y, z), allowing reconstruction of the atomic positions.

If we have two isomorphous (see below) data sets F_{PH} and F_P, then the difference in the two Patterson functions,

$$P_{PH} - P_P = \textstyle\int [F_{PH}^2(\mathbf{S}) - F_P^2(\mathbf{S})] \exp\{-2\pi i \mathbf{r} \cdot \mathbf{S}\} \, \mathrm{d}^3\mathbf{S},$$

will deliver information about the heavy-atom structure. Such a difference function gives rise to non-negligible peaks arising from interference between the F_H and F_P terms, however (Perutz, 1956). Rossmann (1960) showed that these interference terms could be reduced through calculation of the modified Patterson function

$$P_H = \textstyle\int [F_{PH}(\mathbf{S}) - F_P(\mathbf{S})]^2 \exp\{-2\pi i \mathbf{r} \cdot \mathbf{S}\} \, \mathrm{d}^3\mathbf{S}.$$

In the case of a single-site derivative, peaks should occur only at the Harker vectors corresponding to the heavy-atom position. Even so, there is a choice of positions for the heavy atom: *e.g.*, in the $P2_12_12_1$ case, coordinates $(\pm x + \xi, \pm y + \nu, \pm z + \zeta)$, where ξ, ν and ζ can each take the value 0 or 1/2, will all give rise to the same Harker vectors. This in itself is not a problem, relating to equivalent choices of origin and of handedness, but has important ramifications for multisite derivatives or multiple isomorphous replacement (see below).

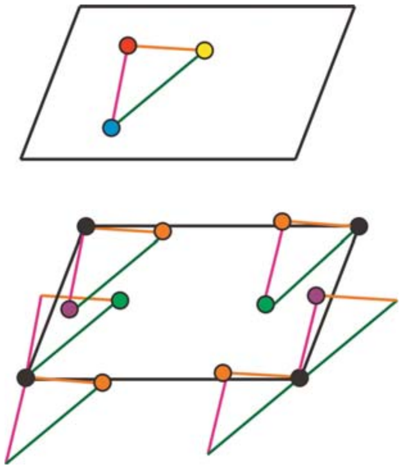


Figure 12.2.2.2
The vector superposition method. The Patterson map of Fig. 12.2.1.1 can be regarded as the superposition of the structure (and its inverse), with each of its atoms placed alternately at the origin. By shifting each peak of the Patterson function to the origin and calculating the correlation of all remaining peaks with the unshifted map, it is possible to deconvolute the Patterson function.

If there is more than one site, then there will be two sets of peaks: one set corresponding to the Harker sections (self-vector set) and one set corresponding to the difference vectors between different heavy-atom sites (the cross-vector set). In this case, the choice of one heavy-atom position (x_{H1}, y_{H1}, z_{H1}) determines the origin and the handedness to which all other peaks *must* correspond. Thus, in the $P2_12_12_1$ example, only one cross vector will occur for

$$(x_{h1} \pm x_{h2} + \xi, y_{h1} \pm y_{h2} + \nu, z_{h1} \pm z_{h2} + \zeta).$$

An alternative to the Harker-vector approach is Patterson-vector superposition (Sheldrick *et al.*, 1993; Richardson & Jacobson, 1987). The Patterson map contains several images of the structure that have been shifted by interatomic vectors (Fig. 12.2.2.2). If this structure is relatively simple (as is to be hoped for in a 'normal' heavy-atom derivative), then it should be possible to deconvolute the superimposed structures by vector shifts (Buerger, 1959).

12.2.3. The difference Fourier

Once the heavy-atom positions have been found, they can be used to calculate approximate phases and Fourier maps. Ideally, difference Fourier maps calculated with phases from a single site should reveal the other positions determined from the Harker search procedure. This ensures that all heavy-atom positions correspond to a single origin and hand. Similarly, phases calculated from derivative $H1$ should reveal the heavy-atom structure for derivative $H2$. Merging and refinement of all phase information will result in a phase set that can be used to solve the structure.

12.2.4. Reality

12.2.4.1. Treatment of errors

Until now, we have dealt with cases involving perfect data. Although this ideal may now be attainable using MAD techniques, this is not necessarily the usual laboratory situation. In the first place, it is necessary to scale the derivative data F_{PH} to the native F_P. One of the most common scaling procedures is based on the expected statistical dependence of intensity on resolution (Wilson, 1949). This may not be particularly accurate when only low-resolution data are available, in which case a scaling through equating the Patterson origin peaks of native and derivative sets may provide better results (Rogers, 1965).

A model to account for errors in the data, determination of heavy-atom positions *etc.* was proposed by Blow & Crick (1959), in which all errors are associated with $|F_{PH}|_{\rm obs}$ (Fig. 12.2.4.1); a more detailed treatment has been provided by Terwilliger & Eisenberg (1987). Owing to errors, the triangle formed by F_P, F_{PH} and F_H fails to close. The *lack of closure error* ε is a function of the calculated phase angle φ_P:

$$\varepsilon(\varphi_P) = |F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}.$$

Once an initial set of heavy-atom positions has been found, it is necessary to refine their parameters (x, y, z, occupancy and thermal parameters). This can be achieved through the minimization of

$$\sum_{\mathbf{S}} \varepsilon^2 / E,$$

where E is the estimated error $(\simeq \langle (|F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc})^2 \rangle)$ (Rossmann, 1960; Terwilliger & Eisenberg, 1983). This procedure is safest for noncentrosymmetric reflections (φ restricted to 0 or π) if enough are present. Phase refinement is generally monitored by three factors:

$$R_{\rm Cullis} = \sum\|F_{PH} + F_P| - |F_H|_{\rm calc}|\big/\sum|F_{PH} - F_P|$$

for noncentrosymmetric reflections only; acceptable values are between 0.4 and 0.6;

$$R_{\rm Kraut} = \sum\||F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}|\big/\sum|F_{PH}|_{\rm obs},$$

which is useful for monitoring convergence; and the

$$\text{phasing power} = \sum|F_H|_{\rm calc}/\sum\||F_{PH}|_{\rm obs} - |F_{PH}|_{\rm calc}|,$$

which should be greater than 1 (if less than 1, then the phase triangle cannot be closed *via* F_H).

The resulting phase probability is given by

$$P(\varphi_P) = \exp\{-\varepsilon^2(\varphi_P)/2E^2\}.$$

The phases have a minimum error when the *best phase* $\varphi_{\rm best}$, *i.e.* the centroid of the phase distribution,

$$\varphi_{\rm best} = \textstyle\int \varphi_P P(\varphi_P)\,{\rm d}\varphi_P,$$

is used instead of the most probable phase. The quality of the phases is indicated by the *figure of merit m*, where

$$m = \textstyle\int P(\varphi_P)\exp(i\varphi_P)\,{\rm d}\varphi_P\big/\int P(\varphi_P)\,{\rm d}\varphi_P.$$

A value of 1 for m indicates no phase error, a value of 0.5 represents a phase error of about 60°, while a value of 0 means that all phases are equally probable.

The *best Fourier* is calculated from

$$\rho_{\rm best}(\mathbf{r}) = (1/V)\sum_{\mathbf{S}} m|F_P(\mathbf{S})|\exp\{i\varphi_{P{\rm best}}(\mathbf{S})\},$$

where the electron density should have minimal errors.

12.2.4.2. Automated search procedures

If the derivative shows a high degree of substitution, then the Harker sections become more difficult to interpret. Furthermore, Terwilliger *et al.* (1987) have shown that the intrinsic noise in the difference Patterson map increases with increasing heavy-atom substitution. It is at this stage that automated procedures are invaluable.

One such automated procedure is implemented in *PROTEIN* (Steigemann, 1991). The unit cell is scanned for possible heavy-atom sites; for each search point (x, y, z), all possible Harker vectors are calculated, and the difference-Patterson-map values at these points are summed or multiplied. As the origin peak dominates the Patterson function, this region is set to zero. The resulting correlation map should contain peaks at all possible heavy-atom positions. The peak list can then be used to find a set of consistent heavy-atom locations through a subsequent search for difference vectors (cross vectors) between putative sites. It should be possible to locate all major and minor heavy-atom sites through repetition of this procedure. A similar strategy is adopted in the program *HEAVY* (Terwilliger *et al.*, 1987), but sets of heavy-atom sites are ranked according to the probability that the peaks are not random. The program *SOLVE* (Terwilliger & Berendzen, 1999) takes this process a stage further, where potential heavy-atom structures are solved and refined to generate an (interpretable) electron density in an automated fashion.

The search method can also be applied in reciprocal space, where the Fourier transform of the trial heavy-atom structure is calculated, and the resulting $F_{H{\rm calc}}$ is compared to the measured differences between derivative and native structure-factor amplitudes (Rossmann *et al.*, 1986). In the programme *XtalView* (McRee, 1998), the correlation coefficient between $|F_H|$ and $|F_{PH} - F_P|$ is calculated, whilst a correlation between F_H^2 and $(F_{PH} - F_P)^2$ is used by Badger & Athay (1998). Dumas (1994*b*,*c*) calculates the correlation between $|F_{H{\rm calc}}|^2$ and $|F_{H{\rm estimated}}|^2$, based on the estimated lack of isomorphism.

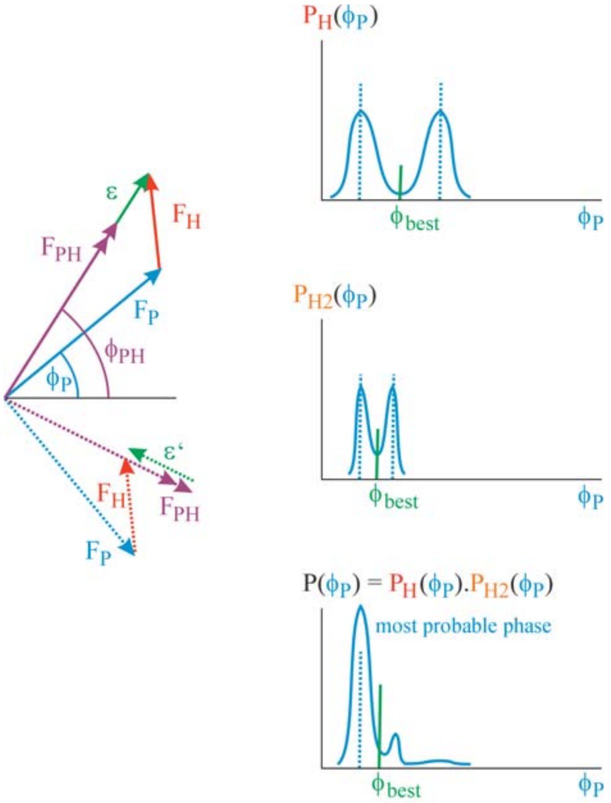


Figure 12.2.4.1

The treatment of phase errors. The calculated heavy-atom structure results in a calculated value for both the phase and magnitude of F_H (red). According to the value of φ_P, the triangle F_P–F_H–F_{PH} will fail to close by an amount ε, the lack of closure (green). This gives rise to a phase distribution which is bimodal for a single derivative. The combined probability from a series of derivatives has a most probable phase (the maximum) and a best phase (the centroid of the distribution), for which the overall phase error is minimum.

Vagin & Teplyakov (1998) have reported a heavy-atom search based on a reciprocal-space translation function. In this case, low-resolution peaks are not removed but weighted down using a Gaussian function. Potential solutions are ranked not only according to their translation-function height, but also through their phasing power, which appears to be a stronger selection criterion.

All these searches are based upon the sequential identification of heavy-atom sites and their incorporation in a heavy-atom partial structure. Problems arise when bogus sites influence the search for further heavy-atom positions. In an attempt to overcome this problem, the heavy-atom search has been reprogrammed using a genetic algorithm, with the Patterson minimum function as a selection criterion (Chang & Lewis, 1994). This approach has the potential to reveal all heavy-atom positions in one calculation, and tests on model data have shown it to be faster than traditional sequential searches.

12.2.5. Special complications

12.2.5.1. Lack of isomorphism

This problem is by far the most common in protein crystallography. An isomorphous derivative is one in which the crystalline arrangement has not been disturbed by derivatization. An early study of Crick & Magdoff (1956) proposed a rule of thumb that a change in any of the cell dimensions by more than around 5% would result in a lack of isomorphism that would defeat any attempt to locate the heavy-atom positions or extract useful phase information. Lack of isomorphism can, however, be more subtle; sometimes a natural variation in the native crystal form may occur, resulting in poor merging statistics of data obtained from different crystals. Coupling this variation with commonly observed structural changes upon heavy-atom binding can provide a considerable barrier to obtaining satisfactory phases. Dumas (1994*a*) has provided a theoretical consideration of this problem.

One practical approach is to collect native and derivative data sets from the same crystal, a technique that has been successful in the structure determination of cyclohydrolase (Nar *et al.*, 1995), proteosome (Löwe *et al.*, 1995) and a number of other proteins. Nonisomorphism can be used, however. In the structure solution of carbamoyl sarcosine hydrolase (Romao *et al.*, 1992), derivatives fell into two (related) crystalline classes. By judicious use of two 'native' crystal forms, heavy-atom positions could be obtained in each of the two classes. Phasing and resultant averaging between the two classes provided an interpretable electron density. In the case of ascorbate oxidase (Messerschmidt *et al.*, 1989), multiple isomorphous replacement failed to provide an interpretable density. It was possible, however, to place the initial density into a second crystal form, which in turn provided phases of sufficient quality to determine heavy-atom sites in derivatives of the second form. Phase-combination and density-modification techniques in the two crystal forms allowed the solution of the structure.

12.2.5.2. Space-group problems

Although the macromolecular crystallographer is rarely confronted with the problems facing their small-molecule colleagues with regard to determining the correct space group, the simplified heavy-atom structure may often throw some surprises. Certain pseudosymmetries may become 'exact' for the heavy-atom difference Patterson map. Thus, cross peaks between different heavy atoms may occur on a Harker section (or 'pseudo-Harker section'), complicating interpretation of the Patterson map. Such was the case with azurin (Adman *et al.*, 1978; Nar *et al.*, 1991), where the heavy-atom structure gave rise to a pseudo-homometric Patterson function, *i.e.* one in which two possible (nonequivalent) choices were available for the heavy-atom structure, only one of which was correct. This arose from a pseudo-centring of the lattice that became almost exact for the heavy-atom structure.

In the case of human NC1 (Stubbs *et al.*, 1990), all heavy-atom derivatives appeared to lie on or near the crystallographic twofold axis. This resulted in a partially centrosymmetric heavy-atom structure that failed to deliver sufficient phase information for noncentrosymmetric reflections. To check for problems with the native data set, anomalous difference Patterson maps {coefficients $[F_{PH}(\mathbf{S}) - F_{PH}(-\mathbf{S})]^2$} were calculated. Coincidence of the peaks obtained from conventional and anomalous Patterson syntheses showed that the heavy-atom positions were correct, but unfortunately did not lead to a structure solution.

12.2.5.3. High levels of substitution; noncrystallographic symmetry

Most problematic are the cases where many heavy atoms have become incorporated in the asymmetric unit. Not only does this cause difficulties in the scaling of derivative to native data, but also the large number of peaks results in ambiguities in the solution of the Patterson function. In such cases, it may be necessary to obtain primary phase information from a different source (such as, for example, another low-substitution-site derivative). One important subclass of high-level substitution is when the native asymmetric unit contains several copies of a single molecule (noncrystallographic symmetry or NCS).

A major problem in locating complex noncrystallographic axes is that the geometrical relationship between NCS peaks in the Patterson map is nontrivial. Under certain conditions, NCS results in a recognizable local symmetry within the Patterson map (Stubbs *et al.*, 1996). In many cases, however, these conditions (that the NCS axes of crystallographic symmetry-related molecules are parallel) are not fulfilled. Under such circumstances, all heavy-atom sites (including all crystallographic symmetry-related positions) must be checked carefully with the rotation function in order to pinpoint the NCS axis. This is relatively trivial for low-order NCS (twofold, threefold), but becomes increasingly complicated for higher orders. It should also always be borne in mind that the heavy-atom positions might not necessarily follow the NCS constraints due to crystal packing. If there is reason to suspect that sites are related by local symmetry, then the orientation of this axis can be used in the initial Harker searches; in practice, however, such searches are extremely sensitive to the correct orientation of the axis.

In the case of high-order NCS (such as, *e.g.*, with icosahedral virus structures or symmetric macromolecular complexes), an alternative approach to the usual initial Harker-vector search can be provided by the self-rotation function. Knowledge of the orientation of the NCS axis (from the rotation function) can be used to determine the relative positions of heavy atoms to the NCS axis (Argos & Rossmann, 1976; Arnold *et al.*, 1987; Tong & Rossmann, 1993). The orientation can be refined and the resulting peaks can be used as input in a subsequent translation search of the Harker sections.

References

Adman, E. T., Stenkamp, R. E., Sieker, L. C. & Jensen, L. H. (1978). *A crystallographic model for azurin at 3.0 Å resolution. J. Mol. Biol.* **123**, 35–47.

Argos, P. & Rossmann, M. G. (1976). *A method to determine heavy-atom positions for virus structures. Acta Cryst.* B**32**, 2975–2983.

Arnold, E., Vriend, G., Luo, M., Griffith, J. P., Kamer, G., Erickson, J. W., Johnson, J. E. & Rossmann, M. G. (1987). *The structure determination of a common cold virus, human rhinovirus 14. Acta Cryst.* A**43**, 346–361.

Badger, J. & Athay, R. (1998). *Automated and graphical methods for locating heavy-atom sites for isomorphous replacement and multiwavelength anomalous diffraction phase determination. J. Appl. Cryst.* **31**, 270–274.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Budisa, N., Karnbrock, W., Steinbacher, S., Humm, A., Prade, L., Neuefeind, T., Moroder, L. & Huber, R. (1997). *Bioincorporation of telluromethionine into proteins: a promising new approach for X-ray structure analysis of proteins. J. Mol. Biol.* **270**, 616–623.

Buerger, M. J. (1959). *Vector Space*. New York: Wiley.

Chang, G. & Lewis, M. (1994). *Using genetic algorithms for solving heavy-atom sites. Acta Cryst.* D**50**, 667–674.

Crick, F. H. C. & Magdoff, B. S. (1956). *The theory of the method of isomorphous replacement for protein crystals. I. Acta Cryst.* **9**, 901–908.

Dumas, P. (1994*a*). *The heavy-atom problem: a statistical analysis. I. A priori determination of best scaling, level of substitution, lack of isomorphism and phasing power. Acta Cryst.* A**50**, 526–537.

Dumas, P. (1994*b*). *The heavy-atom problem: a statistical analysis. II. Consequences of the a priori knowledge of the noise and heavy-atom* *powers and use of a correlation function for heavy-atom-site determination*. *Acta Cryst.* A**50**, 537–546.

Dumas, P. (1994*c*). *The heavy-atom problem: a statistical analysis. II. Consequences of the a priori knowledge of the noise and heavy-atom powers and use of a correlation function for heavy-atom-site determination. Erratum*. *Acta Cryst.* A**50**, 793.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin IV. Sign determination by the isomorphous replacement method*. *Proc. R. Soc. London Ser. A*, **225**, 287–307.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement*. *Acta Cryst.* **9**, 1–9.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure*. *EMBO J.* **9**, 1665–1672.

Löwe, J., Stock, D., Jap, B., Zwickl, P., Baumeister, W. & Huber, R. (1995). *Crystal structure of the 20S proteosome from the archaeon T. acidophilum at 3.4 Å resolution*. *Science*, **268**, 533–539.

McRee, D. E. (1998). *Practical Protein Crystallography*. San Diego: Academic Press.

Messerschmidt, A., Rossi, A., Ladenstein, R., Huber, R., Bolognesi, M., Gatti, G., Marchesini, A., Petruzzelli, R. & Finazzi-Agro, A. (1989). *X-ray crystal structure of the blue oxidase ascorbate oxidase from zucchini. Analysis of the polypeptide fold and a model of the copper sites and ligands*. *J. Mol. Biol.* **206**, 513–529.

Nar, H., Huber, R., Meining, W., Schmid, C., Weinkauf, S. & Bacher, A. (1995). *Atomic structure of GTP cyclohydrolase I*. *Structure*, **3**, 459–466.

Nar, H., Messerschmidt, A., Huber, R., van de Kamp, M. & Canters, G. W. (1991). *X-ray crystal structure of the two site-specific mutants His35Gln and His235Leu of azurin from Pseudomonas aeruginosa*. *J. Mol. Biol.* **218**, 427–447.

Patterson, A. L. (1934). *A Fourier series method for the determination of the components of interatomic distances in crystals*. *Phys. Rev.* **46**, 372–376.

Perutz, M. F. (1956). *Isomorphous replacement and phase determination in non-centrosymmetric space groups*. *Acta Cryst.* **9**, 867–873.

Richardson, J. W. & Jacobson, R. A. (1987). In *Patterson and Pattersons*, edited by J. P. Glusker, B. K. Patterson & M. Rossi. Oxford University Press.

Rogers, D. (1965). In *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 133–148. Oxford University Press.

Romao, M. J., Turk, D., Gomis-Ruth, F. X., Huber, R., Schumacher, G., Mollering, H. & Russmann, L. (1992). *Crystal structure analysis, refinement and enzymatic reaction mechanism of N-carbamoylsarcosine amidohydrolase from Arthrobacter sp. at 2.0 Å resolution*. *J. Mol. Biol.* **226**, 1111–1130.

Rossmann, M. G. (1960). *The accurate determination of the position and shape of heavy-atom replacement groups in proteins*. *Acta Cryst.* **13**, 221–226.

Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method*. New York: Gordon and Breach.

Rossmann, M. G., Arnold, E. & Vriend, G. (1986). *Comparison of vector search and feedback methods for finding heavy-atom sites in isomorphous derivatives*. *Acta Cryst.* A**42**, 325–334.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures*. *Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *The application of direct methods and Patterson interpretation to high-resolution native protein data*. *Acta Cryst.* D**49**, 18–23.

Steigemann, W. (1991). *Recent advances in the PROTEIN program system for the X-ray structure analysis of biological macromolecules.* In *Crystallographic Computing 5: From Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 115–125. Oxford University Press.

Stubbs, M. T., Nar, H., Löwe, J. , Huber, R., Ladenstein, R., Spangfort, M. D. & Svensson, L. A. (1996). *Locating a local symmetry axis from Patterson map cross vectors: application to crystal data from GroEL, GTP cyclohydrolase I and the proteosome*. *Acta Cryst.* D**52**, 447–452.

Stubbs, M. T., Summers, L., Mayr, I., Schneider, M., Bode, W., Huber, R., Ries, A. & Kühn, K. (1990). *Crystals of the NC1 domain of type IV collagen*. *J. Mol. Biol.* **211**, 683–684.

Terwilliger, T. C. & Berendzen, J. (1999). *Automated MAD and MIR structure solution*. *Acta Cryst.* D**55**, 849–861.

Terwilliger, T. C. & Eisenberg, D. (1983). *Unbiased three-dimensional refinement of heavy-atom parameters by correlation of origin-removed Patterson functions*. *Acta Cryst.* A**39**, 813–817.

Terwilliger, T. C. & Eisenberg, D. (1987). *Isomorphous replacement: effects of errors on the phase probability distribution*. *Acta Cryst.* A**43**, 6–13.

Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Generalized method of determining heavy-atom positions using the difference Patterson function*. *Acta Cryst.* A**43**, 1–5.

Tong, L. & Rossmann, M. G. (1993). *Patterson-map interpretation with noncrystallographic symmetry*. *J. Appl. Cryst.* **26**, 15–21.

Vagin, A. & Teplyakov, A. (1998). *A translation-function approach for heavy-atom location in macromolecular crystallography*. *Acta Cryst.* D**54**, 400–402.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities*. *Acta Cryst.* **2**, 318–321.

International Tables for Crystallography (2012). Vol. F, Chapter 13.1, pp. 333–339.

PART 13. MOLECULAR REPLACEMENT

Chapter 13.1. Noncrystallographic symmetry

D. M. Blow

13.1.1. Introduction

Excellent reviews of noncrystallographic symmetry exist. The subject is also discussed in Volume B of this series (Rossmann & Arnold, 2001). Other important reviews include those by Rossmann (1990), Lawrence (1991) and Rossmann (1995). A volume produced for a Daresbury Study Weekend (Dodson *et al.*, 1992) has many interesting chapters.

In this introductory chapter, effort has been made to cite some of the earliest work which initiated the methods which have now become familiar.

13.1.2. Definition of noncrystallographic symmetry

13.1.2.1. Standard noncrystallographic symmetry

The standard cases of noncrystallographic symmetry arise when *there is more than one similar subunit in the crystallographic asymmetric unit*.

The phrase 'noncrystallographic symmetry' is used because the operation required to superimpose one subunit on another is similar to a symmetry operation, but it operates only over a local volume, and the symmetry is inexact because the subunits are in different environments.

The 'subunit' can be a molecular aggregate, a single molecule, a monomer unit of an oligomeric molecule, or a fragment of a molecule.

The word 'similar' is used because protein subunits in different environments are never identical. At the very least, surface side chains are differently ordered, and solvation is different because of different interactions with adjacent subunits.

If noncrystallographic symmetry exists, methods are available to define the operation required to superimpose one unit on another (Rossmann & Blow, 1962; Rossmann *et al.*, 1964). When this has been done, new information is available to improve the accuracy of structural results (Rossmann & Blow, 1963).

Table 13.1.2.1 presents different types of symmetry situations which may arise when noncrystallographic symmetry exists. It is frequently observed that the local symmetry corresponds or approximates to point-group symmetry. This arises very often because a natural molecular form is a symmetric oligomer whose symmetry is not fully expressed in the crystal symmetry [cases (1) and (3)]. Helical symmetry or pseudo-helical symmetry is also common [case (2)], especially in biological materials, but it cannot always be exploited crystallographically because the specimens are often noncrystalline fibres.

13.1.2.2. Generalized noncrystallographic symmetry

Crystallographic methods similar to those which exploit standard noncrystallographic symmetry can often be applied to a more general situation, where *similar subunits exist in different crystals* (Scouloudi, 1969; Tollin, 1969) *or where the structure of a subunit is already predictable* (Hoppe, 1957; Lattman & Love, 1970). The types of relationship which may arise are summarized in the right-hand column of Table 13.1.2.1.

13.1.2.3. Exploitation of noncrystallographic symmetry

In order to draw structural information from noncrystallographic symmetry, the different classes of subunit must provide different information. Cases of pseudo-crystallographic symmetry, where subunits are *almost* in an arrangement of higher crystallographic symmetry, are difficult to exploit by the techniques discussed in this chapter. Typically only weak reflections (those which would be forbidden by the higher symmetry if it were exact) provide extra information. This situation often arises in case (6), Table 13.1.2.1. Similarly in cases (7) and (8), comparison of crystals whose cell dimensions or contents are only slightly altered gives little new information.

13.1.3. Use of the Patterson function to interpret noncrystallographic symmetry

13.1.3.1. Rotation operations

The first step towards identifying and exploiting noncrystallographic symmetry is to find the operation that is required to superimpose one subunit upon another. Superposition of one

Table 13.1.2.1
Noncrystallographic symmetry in crystals

	Relationships within the same crystal	Relationships between different crystals
Symmetry relations	(1) Symmetry of a noncrystallographic point group (*e.g.* 532) (2) Infinite non-closed symmetry (helix) (3) Crystallographic point-group symmetry, not incorporated into lattice	(6) Simple crystallographic relationship between two crystal forms
Relations not forming a group	(4) Similar subunits without systematic relationship	(7) Polymorphism (identical molecules crystallize differently)
Partial structural relationship	(5) Similar subunits account for only a part of unit-cell contents	(8) Crystals of different but similar molecules (9) Crystals containing similar molecules, with other scattering material

Copyright © 2012 International Union of Crystallography

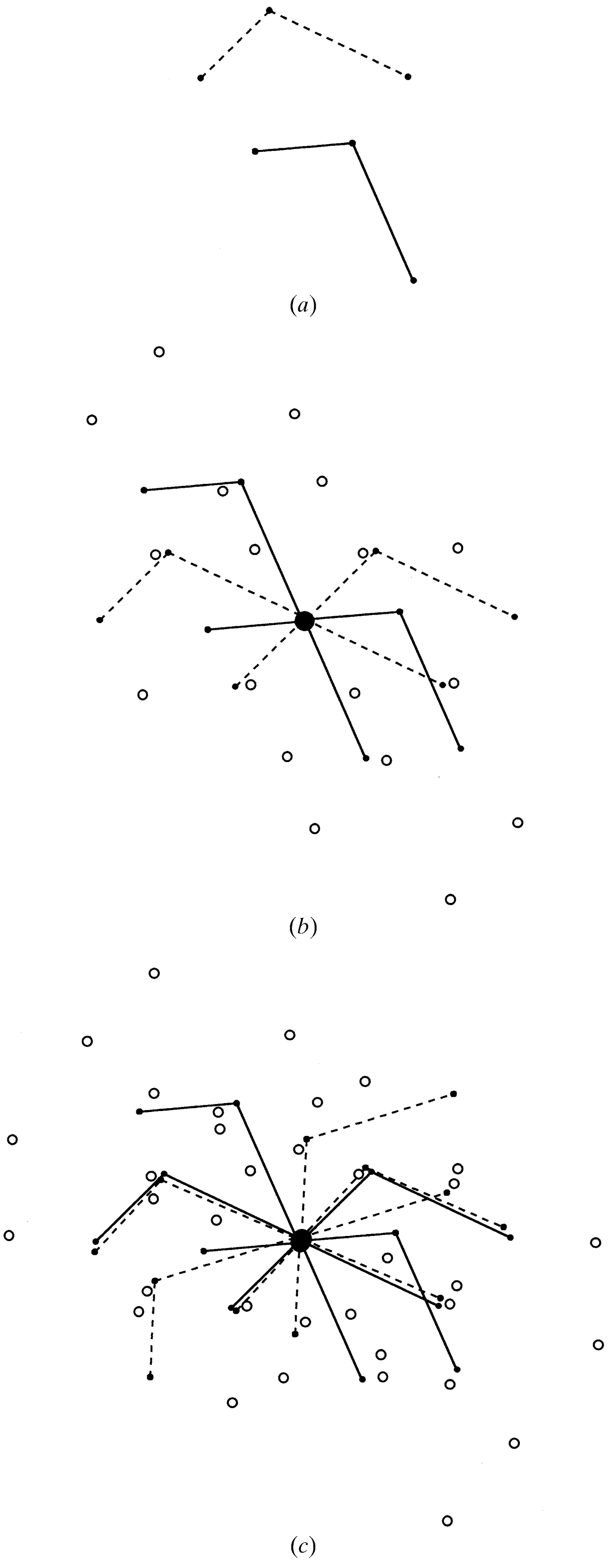


Figure 13.1.3.1
(*a*) A dimer in which the two subunits are related by arbitrary rotation and translation. One member of the dimer is indicated by dashed lines joining its atoms. (*b*) The vector set representing the Patterson function of this dimer. The intra-subunit vectors of each subunit are indicated by filled circles linked by lines; the inter-subunit vectors are shown as open circles. (*c*) A second copy of the Patterson function has been rotated over the original. Intra-subunit vectors of the original Patterson function are indicated by full lines; the rotated intra-subunit vectors are distinguished by dashed lines. The rotation is almost the same as the rotation of one subunit to make it parallel to the other. When the rotation is exactly correct, half of the intra-subunit vectors of one Patterson will superimpose onto the other. None of the inter-subunit vectors superimpose.

asymmetric rigid body upon a similar one requires in general that it be rotated and translated. A general rotation in three-dimensional space requires three variables to specify it: these can be the latitude and longitude of a rotation axis and the angle of rotation (κ, ψ, φ), or they can be the Euler angles. These rotational systems are presented in *IT* B (Rossmann & Arnold, 2001) and are discussed in detail in Chapter 13.2 by Navaza. Similarly, a general three-dimensional translation is specified by three variables. The operation to superimpose subunits therefore requires six variables to define it. Surveys or searches in six-dimensional functions are overwhelmingly laborious, though they have long been possible (Milledge, 1962; Kayushina & Vainshtein, 1965) and are now becoming easier (Kissinger *et al.*, 1999; Sheriff *et al.*, 1999).

The Patterson function, a function directly calculable from observed diffraction intensities, is a function which defines its own origin. Being a function in vector space, its origin is necessarily at a point representing a vector of zero length. It is this special property of the Patterson function which allows its use to factorize the six-dimensional problem into two three-dimensional ones. It is available without information about the phases of the diffraction data. Even when structural data are available, it is usually easier to make a three-dimensional rotation search based on the Patterson function than to carry out a full six-dimensional search using the known structure.

The Patterson function of a crystal may be considered to have two components: vectors between scattering centres in the same subunit, and those between different subunits. The intra-subunit vectors are necessarily shorter than the maximum subunit dimension. Inter-subunit vectors, though a few may be short, are clustered about the distances separating different subunits in the crystal, so they are mostly of the magnitude of the subunit dimensions or longer. By considering the region closer to the origin of the Patterson function, it is possible to include a high proportion of intra-subunit vectors.

Fig. 13.1.3.1 shows how the relation between the set of intra-subunit vectors of one subunit and the intra-subunit vectors of a similar subunit defines a rotation operation. This rotation is identical to the rotational part of the operation required to superimpose the subunits. If the whole Patterson function is rotated in this way, and then superimposed upon itself, one set of intermolecular vectors of the rotated Patterson function is superimposed upon a set of intermolecular vectors of the original (Fig. 13.1.3.1*c*). The self-rotation function (Rossmann & Blow, 1962) searches for correlations between a rotated Patterson function and the original.

By working with Patterson-function vectors, there is no dependence on the relative positions of the two subunits. Relative rotations can be determined in the standard case (Rossmann & Blow, 1962) or in the generalized case (Prothero & Rossmann, 1964; Lattman & Love, 1970). Methods for using the Patterson function to identify the rotation operation are detailed in Chapter 13.2.

13.1.3.2. Translation operations

When the rotation operation has been identified, the translation between the two differently oriented subunits needs to be determined (see Chapter 13.3 by Tong). The translation can only be defined in relation to an assigned origin of the subunit. It is possible to define the translation vector relative to the 'centres of gravity' (more precisely, the centres of scattering density) of the subunits. In this way, a translation may be defined between

Table 13.1.4.1
Structure determination using noncrystallographic symmetry

Starting structural information	Relationships within the same crystal (standard case)	Relationships between different crystals (generalized case)
None	(1) Subunit arrangement defined by relation between noncrystallographic and crystallographic symmetry. Resolution extended by noncrystallographic symmetry constraints	(3)† Subunit arrangement defined by relation between noncrystallographic and crystallographic symmetry in at least one crystal. Cross-rotation and translation functions applied to other crystals. Resolution extended by noncrystallographic symmetry constraints
Poorly resolved structure, unsuitable for refinement	(2) Electron density improved or resolution extended by noncrystallographic symmetry constraints	(4)† Resolution extended by noncrystallographic symmetry constraints
Similar structure known		(5) Subunit orientation found by cross-rotation and translation functions. Phases derived from structural model and may be improved by noncrystallographic symmetry constraints
Part of unknown structure resembles a known structure		(6) Subunit orientation found by cross-rotation and translation functions. Phases derived from structural model and may be improved by noncrystallographic symmetry constraints

† Structure determinations of this kind have not been reported.

subunits of unknown structure. This approach, based only on information in the Patterson function, is the only available method if no phase information is available, but has several difficulties – it only applies accurately in certain cases, and the results are difficult to interpret and are imprecise (Blow *et al.*, 1964; Rossmann *et al.*, 1964). In practice, when dealing with a totally unknown structure, translational relationships are more frequently discovered by using specific markers on the subunit (intense scattering centres or 'heavy atoms', or anomalous scattering centres).

13.1.4. Interpretation of generalized noncrystallographic symmetry where the molecular structure is partially known

13.1.4.1. The cross-rotation function

The rotation function can be used in the generalized case to compare the Patterson functions of different crystals. When used in this way, it is called the cross-rotation function. In the most usual case, information providing some kind of structural model is available for one of the crystals.

The power of the cross-rotation function may be greatly improved by removing all intermolecular vectors from the 'model' Patterson function. This may be done by constructing an imaginary crystal structure in which a single copy of the structural model is placed in a unit cell that is large enough for all intermolecular vectors to be longer than the longest intramolecular vector of the model. In this cell, the self-Patterson vectors may be completely isolated and used for comparison with the Patterson function of a crystal containing a molecule of unknown orientation.

13.1.4.2. The cross-translation function

In searching for the position of a molecule in the generalized case of noncrystallographic symmetry, a molecular model defines an origin of coordinates in the model structure, and the corresponding position can be sought in an unknown structure (Nordman & Nakatsu, 1963; Tollin & Cochran, 1964; Huber, 1965; Crowther & Blow, 1967). The procedure is to calculate a three-dimensional function whose peaks should lie at the intersubunit vectors. In some procedures (Tollin & Cochran, 1964; Crowther & Blow, 1967), this function may be calculated as a Fourier series.

The translation functions will fail if the corresponding rotation is incorrect, or even if it is insufficiently accurate to give a good overlap between the structures. To avoid this danger, Brünger (1997) recommends computing translation functions using rotations corresponding to many (*e.g.* 200) high values of the rotation function. Though this is a huge increase in computing load, it still compares favourably with a full six-dimensional search.

These methods are considered further in Chapter 13.3.

13.1.4.3. Structure determination

Table 13.1.4.1 distinguishes a number of different situations in which noncrystallographic symmetry can be used to aid structure determination. The most frequent application of molecular-replacement methods is to cases where a structure is partially known, but is not yet susceptible to refinement by standard techniques.

Two types of situation arise in the standard case, where noncrystallographically related subunits exist in the same crystal. Most frequently [type (2)], the noncrystallographic symmetry allows the electron density to be improved at the given resolution. Occasionally, high-order noncrystallographic symmetry may be used to extend the resolution to the point where conventional structural refinement becomes possible (Schevitz *et al.*, 1981; McKenna *et al.*, 1992). In the most favourable case, high-order noncrystallographic symmetry constraints may allow direct structure determination [type (1)], starting from the position of a symmetric particle in the asymmetric unit (Jack, 1973).

In the generalized case, most often, similarities with a known molecular structure can be employed to improve an unknown structure [types (5) and (6)]. Such techniques were first used by Tollin (1969) (before structural refinement was possible) and by Fehlhammer & Bode (1975).

It is also possible that a refinable structure could be generated from intensity data observed from several different crystal forms, using noncrystallographic symmetry constraints, but this is not known to have been done in practice [types (3) and (4)].

13.1.5. The power of noncrystallographic symmetry in structure analysis

13.1.5.1. Relevant parameters: standard case

A poorly determined structure, if known at sufficient resolution and accuracy, can be improved by structural refinement of an atomic model to fit the observations. These methods often use existing structural knowledge (of bond lengths and angles, for example) to improve the convergence of the refinement process.

The most important contributions of noncrystallographic symmetry arise before this point of structure determination is reached. In this stage of structural analysis, the distribution of scattering density may be constrained by the requirements of noncrystallographic symmetry. The density may be improved by imposing noncrystallographic symmetry on poorly defined scattering density, and, in favourable cases, cyclical improvement leads to a unique corrected structure. To avoid any confusion with refinement of the atomic structure, this process will be referred to as 'symmetry correction' or 'correction' (Hoppe & Gassmann, 1968).

Noncrystallographic symmetry can also be used to improve the accuracy and convergence of atomic structural refinement by increasing the number of observations to a given resolution (Section 13.1.5.5).

The power of correction methods in improving an unknown structure in the standard case (Section 13.1.2.1) depends on:

(1) the resolution of the analysis, d;
(2) the number, N, of subunits per asymmetric unit;
(3) the volume fraction, NU/V_a, of the asymmetric unit over which the noncrystallographic symmetry operation applies;
(4) whether the density between subunit volumes is constant;
(5) the degree of similarity of the subunits being matched; and
(6) the extent to which the noncrystallographic symmetry operations differ from the crystal symmetry operations.

The first three parameters are expressed in quantitative terms; parameter (4) might be true or false, but more often lies between these; and parameters (5) and (6) are not easily expressed in measurable form.

The resolution d should ideally be matched to the level of similarity of the subunits. The root-mean-square displacement between an atom in one subunit and the rotated and translated position of the corresponding atom from another subunit (or model subunit) provides an order of magnitude for the resolution d which can be used effectively. In many cases, the resolution is worse than this for practical reasons of crystal disorder and data collection.

This limit was encountered by Huber *et al.* (1974) working at 1.9 Å resolution. They found that a model structure with a mean coordinate difference of 1.9 Å was not usable for molecular replacement, while another model agreeing to 0.75 Å gave results which allowed the structure to be refined. This suggests that agreement significantly better than the resolution is required.

13.1.5.2. Information gain from ideal noncrystallographic symmetry

Rossmann & Blow (1962) wrote, "The effect of noncrystallographic symmetry ... results in decreasing the size of the structure to be determined, while the number of observable intensities remains the same. This 'redundancy' in information might be used to help solve a structure." This idea is developed below.

First, it will be shown that (in the absence of noncrystallographic symmetry) there is a constant ratio between the number of independent measurements required to specify the scattering density at a chosen resolution and the volume of the asymmetric unit. Then the effect of noncrystallographic symmetry on this ratio is discussed. The importance of the ratio (volume of symmetry-constrained unit/volume of asymmetric unit) ($= U/V_a$) is stressed. Another ratio is developed – available no. of measurements/ideally required no. of measurements – and this is referred to as the overdetermination ratio.

Consider a noncentrosymmetric crystal whose asymmetric unit volume is V_a and whose diffraction data have been measured to a resolution d. If the multiplicity of the space group (number of asymmetric units in the primitive unit cell) is Z, the volume of reciprocal space, V^*, per point of the primitive reciprocal lattice is given by

$$V^* = 1/V = 1/ZV_a,$$

where V is the volume of the primitive unit cell.

The number of independent orders of diffraction (the number of independent intensities) within the resolution sphere of radius $1/d$ is given by

$$N_{\rm ref} = \left(\frac{4\pi}{3d^3}\right)\left(\frac{1}{V^*}\right)\left(\frac{1}{2Z}\right) = \frac{2\pi V_a}{3d^3}. \qquad (13.1.5.1)$$

In this formula, Friedel's law is supposed to apply. A set of $2Z$ reflections have identical intensity due to the combined effects of Friedel's law and crystal symmetry. When $N_{\rm ref}$ is expressed in terms of V_a, the multiplicity factor disappears.

To calculate the scattering density over the volume V_a at resolution d, $2N_{\rm ref}$ independent quantities need to be specified (say, the real and imaginary parts of each of the $N_{\rm ref}$ independent structure factors). The required number of measurements is

$$R = 2N_{\rm ref} = (4\pi/3d^3)V_a. \qquad (13.1.5.2)$$

If only the diffracted intensities can be measured, they provide exactly half the $2N_{\rm ref}$ measurements required to calculate the density at resolution d.

In what follows, it is assumed that the required number of measurements, R, to specify the scattering density at the chosen resolution is proportional to the volume over which the density must be specified. This is true when the volume is a crystallographic asymmetric unit [equation (13.1.5.2)], and it agrees with another analysis discussed below. Following this argument, the overdetermination ratio

$$\frac{\text{available no. of measurements}}{\text{ideally required no. of measurements}} = \frac{N_{\rm ref}}{R} = \frac{V_a}{2X}, \quad (13.1.5.3)$$

where X is the volume whose density is unknown.

Next, consider that ideal noncrystallographic symmetry applies. The crystal asymmetric unit contains N identical subunits and no other scattering matter. Since the symmetry is noncrystallographic, it is never possible to fit the subunit volumes together so as to fill the unit cell exactly. The volume assigned to each subunit, U, has to be less than V_a/N, leaving some parts of the unit cell not assigned to any subunit. In the case of ideal noncrystallographic symmetry, these regions are necessarily empty. In this case $X = U$, which is less than V_a/N, so from equation (13.1.5.3)

$$\text{overdetermination ratio} = (V_a/2U) > N/2.$$

Even where N is only 2, more intensity data are available than the number of measurements ideally required to specify the electron density at resolution d.

A more sophisticated analysis of the number of variables required to define a structure with noncrystallographic symmetry has been made in terms of sets of orthogonal 'eigendensity functions', which satisfy the noncrystallographic symmetry (Crowther, 1967). Any structure satisfying the symmetry requirements can be constructed from the appropriate set of eigendensities. Crowther (1969) demonstrated that the number of eigendensities m is approximately $(2N_{\rm ref}U/V_a)$.

The structure is specified by m weights, which are applied to the m allowed eigendensities (which depend only on the symmetry constraints), so the overdetermination ratio

$$\frac{\text{available no. of measurements}}{\text{ideally required no. of measurements}} = \frac{N_{\rm ref}}{m} = \frac{V_a}{2U},$$

the same result as before, showing that the two methods of analysis approximately agree.

13.1.5.3. *Information gain in the non-ideal case*

In the non-ideal case, the definition of volume U assigned to each subunit assumes an important role. It is particularly important that the volumes should not overlap, since this may set up a chain of unrealizable constraints. In imposing noncrystallographic symmetry, the volumes between subunits are often unconstrained, allowing for differences in solvent structure and surface side chains following from their different environments. In addition to the volume U of one subunit, whose structure is to be defined, an additional volume, $V_a - NU$, is left unconstrained. The volume X of unknown electron density is $U + (V_a - NU)$, and, using equation (13.1.5.3), the overdetermination ratio

$$\frac{\text{available no. of measurements}}{\text{ideally required no. of measurements}} = \frac{V_a}{2[V_a - (N-1)U]}. \qquad (13.1.5.4)$$

If $N = 2$, U must be less than $V_a/2$, and the overdetermination ratio in equation (13.1.5.4) must be less than 1, so in the non-ideal case there is no chance of convergent correction. This confirms the practical observation that although averaging electron density with $N = 2$ can improve the structure (Matthews *et al.*, 1967), it does not lead to convergent correction (B. W. Matthews, unpublished results). Slowly convergent *ab initio* structure correction was reported at 6.3 Å resolution for $N = 4$ (Argos *et al.*, 1975). In this case, the volume $4U$ of the constrained tetramer was reported to be only about $V_a/2$. Substituting $N = 4$, $U = V_a/8$ in the above expression gives an overdetermination ratio of only 1.6, which was sufficient to allow convergent correction.

An alternative possibility is to constrain the density between subunits to a constant value, even when this may not be precisely correct, in order to improve the convergence of symmetry correction. There is a close analogy to solvent-flattening techniques used in density modification and atomic structural refinement (Schevitz *et al.*, 1981; Wang, 1985). The volume constrained to a constant value is now $V_a - NU$. The volume whose structure is to be determined is only U, and in place of equation (13.1.5.4),

$$\frac{\text{available no. of measurements}}{\text{ideally required no. of measurements}} = \frac{V_a}{2U},$$

as in the ideal case. Such a constraint, while only approximately valid, may allow structure correction to proceed convergently, as found by Rossmann *et al.* (1992). The constraint may be released at a later stage.

This analysis also emphasizes the importance of specifying the size and shape of the subunit volume U as closely as possible (Wilson *et al.*, 1981). Methods of automatic refinement of the chosen volume are available (Rossmann *et al.*, 1992; Abrahams & Leslie, 1996).

13.1.5.4. *Relevant parameters: generalized case*

In the generalized case it is obvious that noncrystallographic symmetry makes more measurements available, since the data from more than one type of crystal are being used. The volume of the subunit U must be less than V_a in each type of crystal. Making a simple assumption of two crystals, each with one subunit in each asymmetric unit, the available number of measurements per volume of the subunit is, in the ideal case,

$$\frac{N_{\rm ref1} + N_{\rm ref2}}{U} = \left(\frac{2\pi}{3d^3}\right)\left(\frac{V_{a1} + V_{a2}}{U}\right) > 2\left(\frac{2\pi}{3d^3}\right).$$

Thus, the overdetermination ratio $= (N_{\rm ref1} + N_{\rm ref2})/(4\pi U/3d^3) > 1$, so even in this case the structure is theoretically overdetermined.

This type of reasoning can be applied to analysis of a crystal which includes a unit of known structure and also another unit whose structure is unknown [type (6), Table 13.1.4.1]. A complex between an enzyme of known structure with an unknown inhibitor provides a familiar example. Note first that the envelope defining the volume U of known structure must be tightly defined, since otherwise unwanted features will be taken over into the unknown structure.

If this known structure of volume U appears once in the asymmetric unit V_a of the partly unknown structure, can noncrystallographic symmetry correction be used to define a unique structure at resolution d?

The unconstrained volume is $V_a - U$. The number of measurements required to define this density at the given resolution is $4\pi(V_a - U)/3d^3$. The partially unknown structure provides $2\pi V_a/3d^3$ measured intensities to this resolution, and specification of the contents of U at resolution d is equivalent to $4\pi U/3d^3$ measurements. Thus the overdetermination ratio is

$$\frac{(2\pi V_a/3d^3) + (4\pi U/3d^3)}{4\pi(V_a - U)/3d^3} = \frac{V_a + 2U}{2(V_a - U)}.$$

The overdetermination ratio is greater than 1 if $U > V_a/4$, that is, if only a quarter of the asymmetric unit represents known structure. Although this relationship applies to an 'ideal' case and is therefore certainly too optimistic, it indicates the remarkable power of the molecular-replacement method. If, for example, half the unit cell is devoted to unknown structure ($U = V_a/2$), the overdetermination ratio is ideally 2.

13.1.5.5. *Noncrystallographic symmetry in atomic coordinate refinement*

In atomic coordinate refinement, noncrystallographic symmetry again provides a useful increase in the ratio of the number of observed quantities to the number of atomic parameters to be refined. As is discussed by Cruickshank in Chapter 18.5, the application of restraints in refinement (on quantities like bond lengths, bond angles and the elimination of short contacts) is formally equivalent to an increase in the number of observational equations. However, if these restraints are tightly applied, they act more like constraints, and their effect is more like a reduction in the number of parameters to be determined. Meaningful refinement is not possible unless the number of observations exceeds the number of parameters, and in practice it usually needs to do so by a factor of 2 or so. If noncrystallographic symmetry is imposed, the number of observations required to define the structure is reduced, because the volume of unknown structure is reduced.

Noncrystallographic symmetry can thus provide a crucial advantage in leading to unambiguous interpretation of structure at relatively poor resolution (say, 3.0 to 3.8 Å), where the ratio of refined parameters to the number of observations is marginal. Consider two crystals of the same material, one of which has one subunit per asymmetric unit and the asymmetric volume is V_1. The other has N subunits in an asymmetric unit of volume V_N. To the same resolution, the available number of observed reflections is increased in the ratio V_N/V_1, in order to obtain the same number of parameters if noncrystallographic symmetry is imposed.

What effect do these N subunits have on the precision of the final coordinates? The crystal allows the determination of N sets of atomic coordinates. If the errors were independent of each other, the precision of the mean value of each coordinate could be improved in the ratio $N^{-1/2}$ (compared to a well refined V_1 structure).

This improvement will be lost when constraints are applied to the mean coordinates (to make them conform to given bond lengths and angles, for example). If this is done, the errors are no longer independent, and the increase of precision will be less.

Cruickshank (1999*a*,*b* and Chapter 18.5) shows that at high resolution (examples at 0.94 and 1.0 Å) and for atoms of low B factor (less than say 10 Å^2), restraints make little difference to the precision of refinement. Under these conditions, N independent subunits in the asymmetric unit might improve the precision of the mean coordinates by a factor approaching $N^{-1/2}$. But at such good resolution, it is very possible that the differences between the calculated subunit conformations are not due to error, but reflect real structural differences. If so, the precision of the mean coordinates is less significant.

At less high resolution (example given at 1.7 Å), Cruickshank has shown that the precision of unrestrained refinement is significantly worse than the precision of the restraints. In this case, imposing noncrystallographic symmetry on the structure should provide some improvement. But because the coordinate errors then cease to be independent, the improvement in the mean coordinates would be less.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F_1 ATPase. Acta Cryst.* D**52**, 30–42.

Argos, P., Ford, G. C. & Rossmann, M. G. (1975). *An application of the molecular replacement technique in direct space to a known protein structure. Acta Cryst.* A**31**, 499–506.

Blow, D. M., Rossmann, M. G. & Jeffery, B. A. (1964). *The arrangement of α-chymotrypsin molecules in the monoclinic crystal form. J. Mol. Biol.* **8**, 65–78.

Brünger, A. T. (1997). *Patterson correlation searches and refinement. Methods Enzymol.* **276**, 558–580.

Crowther, R. A. (1967). *A linear analysis of the non-crystallographic symmetry problem. Acta Cryst.* **22**, 758–764.

Crowther, R. A. (1969). *The use of non-crystallographic symmetry for phase determination. Acta Cryst.* B**25**, 2571–2580.

Crowther, R. A. & Blow, D. M. (1967). *A method of positioning a known molecule in an unknown crystal structure. Acta Cryst.* **23**, 544–548.

Cruickshank, D. W. J. (1999*a*). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (1999*b*). *Remarks about protein structure precision. Erratum. Acta Cryst.* D**55**, 1108.

Dodson, E., Gover, S. & Wolf, W. (1992). Editors. *Proceedings of the CCP4 Study Weekend. Molecular Replacement.* Warrington: Daresbury Laboratory.

Fehlhammer, H. & Bode, W. (1975). *The refined crystal structure of bovine β-trypsin at 1.8 Å resolution. I. Crystallization, data collection and application of Patterson search techniques. J. Mol. Biol.* **98**, 683–692.

Hoppe, W. (1957). *Die 'Faltmolekülmethode' – eine neue Methode zur Bestimmung der Kristallstruktur bei ganz oder teilweise bekannter Molekülstruktur. Acta Cryst.* **10**, 750–751.

Hoppe, W. & Gassmann, J. (1968). *Phase correction, a new method to solve partially known structures. Acta Cryst.* B**24**, 97–107.

Huber, R. (1965). *Die automatisierte Faltmolekülmethode. Acta Cryst.* **19**, 353–356.

Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J. & Steigemann, W. (1974). *Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II. Crystallographic refinement at 1.9 Å resolution. J. Mol. Biol.* **89**, 73–101.

Jack, A. (1973). *Direct determination of X-ray phases for tobacco mosaic virus protein using non-crystallographic symmetry. Acta Cryst.* A**29**, 545–554.

Kayushina, R. L. & Vainshtein, B. K. (1965). *Rentgenografiye opredepenie strukturi L-prolina. Kristallografiya*, **10**, 833–844.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Rapid automated molecular replacement by evolutionary search. Acta Cryst.* D**55**, 484–491.

Lattman, E. E. & Love, W. E. (1970). *A rotational search procedure for detecting a known molecule in a crystal. Acta Cryst.* B**26**, 1854–1857.

Lawrence, M. C. (1991). *The application of the molecular replacement method to the de novo determination of protein structure. Q. Rev. Biophys.* **24**, 399–424.

McKenna, R., Xia, D., Willingmann, P., Ilag, L. L. & Rossmann, M. G. (1992). *Structure determination of the bacteriophage φX174. Acta Cryst.* B**48**, 499–511.

Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Three-dimensional structure of tosyl-α-chymotrypsin. Nature (London)*, **214**, 652–656.

Milledge, H. J. (1962). *The automatic selection of molecular-crystal structure by combining stereochemical criteria and high-speed computing. Proc. R. Soc. London Ser. A*, **267**, 566–589.

Nordman, C. E. & Nakatsu, K. (1963). *Interpretation of the Patterson function of crystals containing a known molecular fragment. The structure of an Alstonia alkaloid. J. Am. Chem. Soc.* **85**, 353.

Prothero, J. W. & Rossmann, M. G. (1964). *The relative orientation of molecules of crystallized human and horse oxyhaemoglobin. Acta Cryst.* **17**, 768–769.

Rossmann, M. G. (1990). *The molecular replacement method. Acta Cryst.* A**46**, 73–82.

Rossmann, M. G. (1995). *Ab initio phase determination and phase extension using non-crystallographic symmetry. Curr. Opin. Struct. Biol.* **5**, 650–655.

Rossmann, M. G. & Arnold, E. (2001). *Patterson and molecular-replacement techniques.* In *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.3. Dordrecht: Kluwer Academic Publishers.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Rossmann, M. G. & Blow, D. M. (1963). *Determination of phases by the conditions of non-crystallographic symmetry. Acta Cryst.* **16**, 39–45.

Rossmann, M. G., Blow, D. M., Harding, M. M. & Coller, E. (1964). *The relative positions of independent molecules within the same asymmetric unit. Acta Cryst.* **17**, 338–342.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *Molecular replacement real-space averaging. J. Appl. Cryst.* **25**, 166–180.

Schevitz, R. W., Podjarny, A. D., Zwick, M., Hughes, J. J. & Sigler, P. B. (1981). *Improving and extending the phases of medium- and low-resolution macromolecular structure factors by density modification. Acta Cryst.* A**37**, 669–677.

Scouloudi, H. (1969). *X-ray crystallographic studies of seal myoglobin at 6-Å and 5-Å resolution. J. Mol. Biol.* **40**, 353–377.

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *Implementation of a six-dimensional search using the AMoRe translation function for difficult molecular-replacement problems. J. Appl. Cryst.* **32**, 98–101.

Tollin, P. (1969). *Determination of the orientation and position of the myoglobin molecule in the crystal of seal myoglobin. J. Mol. Biol.* **45**, 481–490.

Tollin, P. & Cochran, W. (1964). *Patterson function interpretation for molecules containing planar groups. Acta Cryst.* **17**, 1322–1324.

Wang, B. C. (1985). *Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol.* **115**, 90–92.

Wilson, I. A., Skehel, J. J. & Wiley, D. C. (1981). *Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature (London)*, **289**, 366–373.

Chapter 13.2. Rotation functions

J. Navaza

13.2.1. Overview

We will discuss a technique to find either the relative orientations of homologous but independent subunits connected by noncrystallographic symmetry (NCS) elements or the absolute orientations of these subunits if the structure of a similar molecule or fragment is available. The procedure makes intensive use of properties of the rotation group, so we will start by recalling some properties of rotations. More advanced results are included in Appendix 13.2.1.

13.2.2. Rotations in three-dimensional Euclidean space

A rotation **R** is specified by an oriented axis, characterized by the unit vector **u**, and the spin, χ, about it. Positive spins are defined by the right-hand screw sense and values are given in degrees. An almost one-to-one correspondence between rotations and parameters $(\chi, \mathbf{u})$ can be established. If we restrict the spin values to the positive interval $0 \leq \chi \leq 180$, then for each rotation there is a unique vector $\chi\mathbf{u}$ within the sphere of radius 180. However, vectors situated at opposite points on the surface correspond to the same rotation, *e.g.* $(180, \mathbf{u})$ and $(180, -\mathbf{u})$.

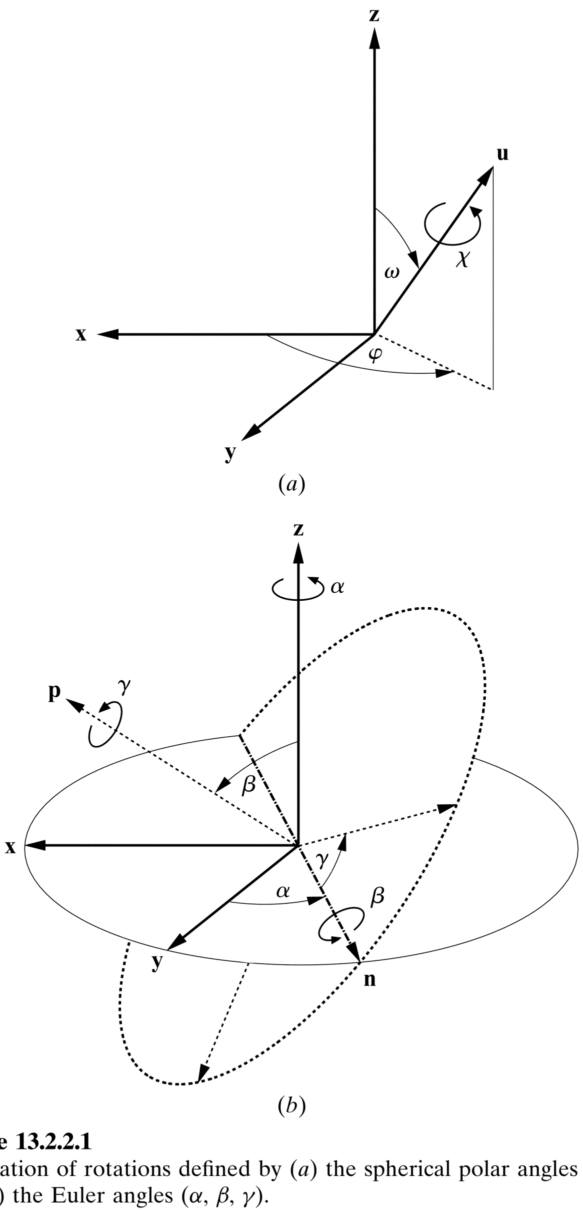


Figure 13.2.2.1
Illustration of rotations defined by (*a*) the spherical polar angles (χ, ω, φ); (*b*) the Euler angles (α, β, γ).

When the unit vector **u** is specified by the colatitude ω and the longitude φ with respect to an orthonormal reference frame (see Fig. 13.2.2.1*a*), we have the spherical polar parameterization of rotations (χ, ω, φ). The range of variation of the parameters is

$$0 \leq \chi \leq 180;\ \ 0 \leq \omega \leq 180;\ \ 0 \leq \varphi < 360.$$

Rotations may also be parameterized with the Euler angles (α, β, γ) associated with an orthonormal frame $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. Several conventions exist for the names of angles and definitions of the axes involved in this parameterization. We will follow the convention by which (α, β, γ) denotes a rotation of α about the z axis, followed by a rotation of β about the nodal line n, the rotated y axis, and finally a rotation of γ about **p**, the rotated z axis (see Fig. 13.2.2.1*b*):

$$\mathbf{R}(\alpha, \beta, \gamma) = \mathbf{R}(\gamma, \mathbf{p})\mathbf{R}(\beta, \mathbf{n})\mathbf{R}(\alpha, \mathbf{z}). \tag{13.2.2.1}$$

The same rotation may be written in terms of rotations around the fixed orthonormal axes. By using the group property

$$\mathbf{T}\mathbf{R}(\chi, \mathbf{u})\mathbf{T}^{-1} = \mathbf{R}(\chi, \mathbf{T}\mathbf{u}), \tag{13.2.2.2}$$

which is valid for any rotation **T**, we obtain (see Appendix 13.2.1)

$$\mathbf{R}(\alpha, \beta, \gamma) = \mathbf{R}(\alpha, \mathbf{z})\mathbf{R}(\beta, \mathbf{y})\mathbf{R}(\gamma, \mathbf{z}). \tag{13.2.2.3}$$

The parameters (α, β, γ) take values within the parallelepiped

$$0 \leq \alpha < 360;\ \ 0 \leq \beta \leq 180;\ \ 0 \leq \gamma < 360.$$

Here again, different values of the parameters may correspond to the same rotation, *e.g.* $(\alpha, 180, \gamma)$ and $(\alpha - \gamma, 180, 0)$.

Although rotations are abstract objects, there is a one-to-one correspondence with the orthogonal matrices in three-dimensional space. In the following sections, **R** will denote a 3×3 orthogonal matrix. An explicit expression for the matrix which corresponds to the rotation $(\chi, \mathbf{u})$ is

$$\begin{bmatrix} \cos\chi + u_1u_1(1-\cos\chi) & u_1u_2(1-\cos\chi) - u_3\sin\chi & u_1u_3(1-\cos\chi) + u_2\sin\chi \\ u_2u_1(1-\cos\chi) + u_3\sin\chi & \cos\chi + u_2u_2(1-\cos\chi) & u_2u_3(1-\cos\chi) - u_1\sin\chi \\ u_3u_1(1-\cos\chi) - u_2\sin\chi & u_3u_2(1-\cos\chi) + u_1\sin\chi & \cos\chi + u_3u_3(1-\cos\chi) \end{bmatrix} \tag{13.2.2.4}$$

or, in condensed form,

$$\mathbf{R}(\chi, \mathbf{u})_{ij} = \delta_{ij}\cos\chi + u_iu_j(1-\cos\chi) + \sum_{k=1}^{3}\varepsilon_{ikj}u_k\sin\chi, \tag{13.2.2.5}$$

where δ_{ij} is the Kronecker tensor, u_i are the components of **u**, and ε_{ijk} is the Levi–Civita tensor. The rotation matrix in the Euler parameterization is obtained by substituting the matrices in the right-hand side of equation (13.2.2.3) by the corresponding expressions given by equation (13.2.2.4).

13.2.2.1. The metric of the rotation group

The idea of distance between rotations is necessary for a correct formulation of the problem of sampling and for plotting functions of rotations (Burdina, 1971; Lattman, 1972). It can be demonstrated that the quantity

$$\mathrm{d}s^2 = \mathrm{Tr}\big(\mathrm{d}\mathbf{R}\ \mathrm{d}\mathbf{R}^{+}\big) = \sum_{i,j=1}^{3} (\mathrm{d}R_{ij})^2 \qquad (13.2.2.6)$$

defines a metric on the rotation group, unique up to a multiplicative constant, which cannot be reduced to a Cartesian metric. This is a topological property of the group, independent of its parameterization. ds is interpreted as the distance between the rotations $\mathbf{R}$ and $\mathbf{R} + \mathrm{d}\mathbf{R}$. With the Euler parameterization, equation (13.2.2.6) becomes

$$\mathrm{d}s^2 = \mathrm{d}\alpha^2 + 2\cos(\beta)\ \mathrm{d}\alpha\ \mathrm{d}\gamma + \mathrm{d}\gamma^2 + \mathrm{d}\beta^2. \qquad (13.2.2.7)$$

The volume element of integration, $\sin(\beta)\ \mathrm{d}\alpha\ \mathrm{d}\beta\ \mathrm{d}\gamma$, corresponding to this length element guarantees the invariance of integrals over the rotation angles with respect to initial reference orientations.

The distance defined by equation (13.2.2.6) has a simple physical interpretation. Let us consider a molecule with initial atomic coordinates $\{\mathbf{x}\}$, referred to an orthonormal frame parallel to the molecule's principal moments of inertia, I_i. Then, the coordinates satisfy the conditions

$$\begin{aligned} \langle x_i\ x_j\rangle &= 0 \quad \text{if } i \neq j,\\ \langle x_i\ x_i\rangle &= I_i, \end{aligned} \qquad (13.2.2.8)$$

where $\langle \ldots \rangle$ means 'average over atoms'. If we move the molecule from a rotated position characterized by the rotation $\mathbf{R}$ to a close one characterized by $\mathbf{R} + \mathrm{d}\mathbf{R}$, the mean-square shift of the atomic coordinates is

$$\sigma^2 = \langle (\mathrm{d}\mathbf{R}\ \mathbf{x})^2\rangle = \sum_{i=1}^{3} I_i \sum_{j=1}^{3} (\mathrm{d}R_{ij})^2. \qquad (13.2.2.9)$$

When all the I_i are equal, σ becomes proportional to ds.

13.2.3. The rotation function

The absolute and relative orientations of the subunits that constitute a crystal may in principle be found by exploiting some properties of the Patterson function. We first express the crystal structure factors $F(\mathbf{h})$ in terms of the Fourier transform of the electron density of isolated molecules, $f_m(\mathbf{s})$, calculated with their centres of gravity placed at the origin. If $\mathbf{r}_m$ denotes the position within the crystal of the centre of gravity of the mth molecule, we have

$$F(\mathbf{h}) = \sum_{m=1}^{M} f_m(\mathbf{h}) \exp(2\pi i \mathbf{h}\mathbf{r}_m), \qquad (13.2.3.1)$$

where M is the number of molecules within the unit cell. This is a general expression for $F(\mathbf{h})$, although the space-group and noncrystallographic rotational symmetries are not explicitly exhibited. The presence of symmetry implies that some of the $f_m(\mathbf{h})$'s are in fact samples of the same function at rotated arguments. The Fourier coefficients of the Patterson function can then be written as

$$\begin{aligned} I(\mathbf{h}) = |F(\mathbf{h})|^2 &= \sum_{m=1}^{M} |f_m(\mathbf{h})|^2 \\ &+ \sum_{m\neq m'=1}^{M} \overline{f_m(\mathbf{h})}\, f_{m'}(\mathbf{h}) \exp[2\pi i \mathbf{h}(\mathbf{r}_m - \mathbf{r}_{m'})], \end{aligned} \qquad (13.2.3.2)$$

where $\overline{f_m}$ is the complex conjugate of f_m. The first term in equation (13.2.3.2) involves only intramolecular contributions or self-Patterson terms; they are centred at the origin of the Patterson function. The second term involves intermolecular contributions or cross-Patterson terms; they are centred at the intermolecular vectors $\mathbf{r}_m - \mathbf{r}_{m'}$. Therefore, if we restrict the Patterson function to a region Ω, of volume v, centred at the origin and having a dimension of the order of the dimensions of the isolated molecules, self-Patterson terms will dominate the crossed ones.

Thus it becomes possible to determine the rotation $\mathbf{R}$ that superimposes one subunit upon an independent, homologous one by calculating the overlap within the region Ω of the observed Patterson function (the target function P_t) and a rotated version of either itself or the Patterson function of an isolated known molecule (the search function P_s):

$$\mathcal{R}(\mathbf{R}) = (1/v) \int_{\Omega} P_t(\mathbf{r}) P_s(\mathbf{R}^{-1}\mathbf{r})\ \mathrm{d}^3\mathbf{r} \qquad (13.2.3.3)$$

(Rossmann & Blow, 1962). $\mathcal{R}$ should display a local maximum for the sought rotation. Note that when we rotate the search function P_s by $\mathbf{R}$, its argument contains $\mathbf{R}^{-1}$. When the search and the target functions are the same, $\mathcal{R}$ is called the self-rotation function; otherwise, it is called the cross-rotation function. When a model is available, the target data are calculated by placing the model within a $P1$ cell whose dimensions guarantee that Ω only contains self-Patterson contributions.

The reciprocal-space formulation of the above integral is obtained by substituting the Patterson functions by their Fourier summations:

$$P(\mathbf{r}) = \sum_{\mathbf{h}} [I(\mathbf{h})/V] \exp(-2\pi i \mathbf{h}\mathbf{r}). \qquad (13.2.3.4)$$

Taking into account that $I(-\mathbf{h}) = I(\mathbf{h})$, we obtain

$$\begin{aligned} \mathcal{R}(\mathbf{R}) &= \sum_{\mathbf{h}}\sum_{\mathbf{k}} \frac{I_t(\mathbf{h})}{V_t}\frac{I_s(\mathbf{k})}{V_s}\frac{1}{v}\int_{\Omega} \exp[2\pi i(\mathbf{h} - \mathbf{k}\mathbf{R}^{-1})\mathbf{r}]\ \mathrm{d}^3\mathbf{r} \\ &= \sum_{\mathbf{h}}\sum_{\mathbf{k}} \frac{I_t(\mathbf{h})}{V_t}\frac{I_s(\mathbf{k})}{V_s}\ \chi_{\Omega}(\mathbf{h} - \mathbf{k}\mathbf{R}^{-1}). \end{aligned} \qquad (13.2.3.5)$$

χ_{Ω}, the interference function, is the Fourier transform of the characteristic function of Ω, *i.e.*, a function that takes the value 1 within Ω and 0 outside. In principle, the domain of integration could have any shape. However, in order to take full advantage of the properties of the rotation group, Ω is usually chosen as a spherical domain of radius b. Calling $\mathbf{h} - \mathbf{k}\mathbf{R}^{-1} = \mathbf{s}$ for short, we have

$$\begin{aligned} \chi_b(\mathbf{s}) &= (3/4\pi b^3) \int_0^b\int_0^{\pi}\int_0^{2\pi} \exp(2\pi i \mathbf{s}\mathbf{r}) r^2 \sin(\theta)\ \mathrm{d}r\ \mathrm{d}\theta\ \mathrm{d}\varphi \\ &= 3[\sin(2\pi sb) - 2\pi sb\cos(2\pi sb)]\big/(2\pi sb)^3. \end{aligned} \qquad (13.2.3.6)$$

In the case of a spherical shell with inner and outer radii a and b, respectively, the interference function is obtained by subtraction:

$$[\chi_b(\mathbf{s}) - (a/b)^3 \chi_a(\mathbf{s})]\big/[1 - (a/b)^3]. \qquad (13.2.3.7)$$

Although simple, the resulting expression for the rotation function has the disadvantage of containing entangled $\mathbf{h}$, $\mathbf{k}$ and $\mathbf{R}$ contributions, which renders its computation time-consuming if the whole domain of rotations has to be explored. The difficulty may be overcome by expanding the exponentials entering equation (13.2.3.5) in spherical harmonics, $Y_{\ell,m}$. Taking advantage of their transformation under rotations, and using recurrence relationships between spherical Bessel functions, j_{ℓ}, we obtain (see Appendix 13.2.1)

$$\chi_b(\mathbf{h}-\mathbf{k}\mathbf{R}^{-1}) = \sum_{\ell=0}^{\infty}\sum_{m,\,m'=-\ell}^{\ell} \overline{Y_{\ell,\,m}(\hat{\mathbf{h}})}\; Y_{\ell,\,m'}(\hat{\mathbf{k}}) \times \left\{\sum_{n=1}^{\infty} 12\pi[2(\ell+2n)-1]\,\frac{j_{\ell+2n-1}(2\pi hb)}{2\pi hb}\,\frac{j_{\ell+2n-1}(2\pi kb)}{2\pi kb}\right\} \times D^{\ell}_{m,\,m'}(\mathbf{R}), \tag{13.2.3.8}$$

where $D^{\ell}_{m,\,m'}$ are the matrices of the irreducible representations of the rotation group and $\hat{\mathbf{h}}$ stands for 'angular part of vector $\mathbf{h}$'. The awkwardness of equation (13.2.3.8) is apparent rather than real. Indeed:

(1) The expression separates angular from crystal variables. It also separates target from search contributions (see Section 13.2.3.1).

(2) The equation is accurate, even when truncating the summations on ℓ and n to reasonable values (Navaza, 1993). The upper limit for ℓ is of the order of the highest argument of the spherical Bessel functions,

$$\ell_{\max} \simeq 2\pi b/d_{\min},$$

where $d_{\min}$ is the resolution of the data. The upper limit for n also depends on the current value of ℓ,

$$n_{\max}(\ell) \simeq (\ell_{\max} - \ell + 2)/2.$$

(3) When the rotations are parameterized in Euler angles (α, β, γ), the matrices $D^{\ell}_{m,\,m'}$ take the form

$$D^{\ell}_{m,\,m'}(\alpha,\beta,\gamma) = d^{\ell}_{m,\,m'}(\beta)\exp[i(m\alpha + m'\gamma)], \tag{13.2.3.9}$$

which enables the computation of $\mathcal{R}$, for each given value of β, by means of two-dimensional fast Fourier transforms (FFTs).

This formulation is referred to as the fast rotation function (Crowther, 1972).

It may be useful to compare rotation functions obtained under different conditions. For this, some kind of normalization is needed. One possibility is to cast $\mathcal{R}$ into the form of a correlation coefficient by dividing equation (13.2.3.3) by the norms of the truncated Patterson functions,

$$\mathcal{R}_N(\mathbf{R}) = \int_{\Omega} P_t(\mathbf{r})P_s(\mathbf{R}^{-1}\mathbf{r})\,\mathrm{d}^3\mathbf{r} \bigg/ \left[\int_{\Omega} P_t(\mathbf{r})^2\,\mathrm{d}^3\mathbf{r} \int_{\Omega} P_s(\mathbf{r})^2\,\mathrm{d}^3\mathbf{r}\right]^{1/2}. \tag{13.2.3.10}$$

13.2.3.1. *Computing the rotation function*

A direct evaluation of equation (13.2.3.3) is possible. However, since the values of the Patterson functions are only available at discrete sampling points, an interpolation is needed after each rotation. This method is known as the direct-space formulation of the rotation function (Nordman, 1966; Steigemann, 1974).

In the reciprocal-space formulation [equations (13.2.3.5) and (13.2.3.6)], some precautions are to be taken in order to save computing time:

(1) Use only strong reflections in the summations.

(2) For each $\mathbf{k}\mathbf{R}^{-1}$, limit the summation on $\mathbf{h}$ to vectors satisfying the condition $|\mathbf{h} - \mathbf{k}\mathbf{R}^{-1}| \leq 1/2\pi b$.

In both formulations, it is worth using a distance between rotations to sample $\mathcal{R}$ efficiently, as explained in Section 13.2.3.2.

For the fast rotation function [equations (13.2.3.5) and (13.2.3.8)], the calculations are organized as follows (Dodson, 1985; Navaza, 1993):

(1) Given the search and the target diffraction data, compute

$$e_{\ell,\,m,\,n} = \left\{12\pi[2(\ell+2n)-1]\right\}^{1/2} \times \sum_{\mathbf{h}} [I(\mathbf{h})/V]\,Y_{\ell,\,m}(\hat{\mathbf{h}})[j_{\ell+2n-1}(2\pi hb)]/2\pi hb \tag{13.2.3.11}$$

and normalize [to compute equation (13.2.3.10)]:

$$e_{\ell,\,m,\,n} \rightarrow e_{\ell,\,m,\,n} \bigg/ \left(\sum_{\ell=2}^{\ell_{\max}} \sum_{m=-\ell}^{\ell} \sum_{n=1}^{n_{\max}(\ell)} |e_{\ell,\,m,\,n}|^2\right)^{1/2}. \tag{13.2.3.12}$$

Odd ℓ terms disappear, because the Friedel-related reflections contribute with opposite signs:

$$I(-\mathbf{h})Y_{\ell,\,m}(-\hat{\mathbf{h}}) = (-1)^{\ell} I(\mathbf{h})Y_{\ell,\,m}(\hat{\mathbf{h}}).$$

Also, if the Patterson function has an n-fold rotation axis along $\mathbf{z}$, only the terms with m equal to a multiple of n survive.

(2) Given the $e_{\ell,\,m,\,n}$'s, perform the sums

$$C^{\ell}_{m,\,m'} = \sum_{n=1}^{n_{\max}(\ell)} \overline{e^{(t)}_{\ell,\,m,\,n}}\; e^{(s)}_{\ell,\,m',\,n}. \tag{13.2.3.13}$$

(3) For each β value, calculate the reduced matrix elements and compute

$$S_{m,\,m'}(\beta) = \sum_{\ell=2}^{\ell_{\max}} C^{\ell}_{m,\,m'} d^{\ell}_{m,\,m'}(\beta). \tag{13.2.3.14}$$

Then evaluate the β section of $\mathcal{R}$ by FFT:

$$\mathcal{R}(\alpha,\beta,\gamma) = \sum_{m,\,m'=-\ell_{\max}}^{\ell_{\max}} S_{m,\,m'}(\beta)\exp[i(m\alpha + m'\gamma)]. \tag{13.2.3.15}$$

The sampling in α and γ here is dictated by the standard FFT requirements.

13.2.3.2. *Plotting and sampling the rotation function*

An odd feature of the rotation domain is that quite different values of the parameters may correspond to very close rotations. This causes the graphic representation of the rotation function to be distorted. For example, when the Euler angle β is small or close to 180, a peak plotted on a Cartesian grid in α and γ is strongly elongated in one direction. By redefining the reference orientation, the same peak near the $\beta = 90$ section will have a quite different – natural – shape.

This problem may be partially overcome by using the distance between rotations introduced in Section 13.2.2.1. Although the metric cannot be reduced to a Cartesian one in three dimensions, it is possible to do so for two-dimensional sections. Indeed, for fixed β, the transformation

$$\begin{aligned} \omega_+ &= \cos(\beta/2)(\alpha+\gamma) \\ \omega_- &= \sin(\beta/2)(\alpha-\gamma) \end{aligned} \tag{13.2.3.16}$$

reduces equation (13.2.2.7) to $\mathrm{d}s^2 = \mathrm{d}\omega_+^2 + \mathrm{d}\omega_-^2$. Therefore, distortion-free β sections may be plotted in terms of the variables $\omega_{\pm}$ (Burdina, 1971; Lattman, 1972). Accordingly, a correct sampling of $\mathcal{R}$ should consist of points regularly spaced in these variables. The number of points with respect to that of a regular spacing in α, β and γ is approximately

$$\int \sin(\beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta\,\mathrm{d}\gamma \Big/ \int \mathrm{d}\alpha\,\mathrm{d}\beta\,\mathrm{d}\gamma = 2/\pi \simeq 0.64.$$

Notice that for β equal to 0 or 180, the sections reduce to lines.

In the case of the self-rotation function, when the order of the NCS may be anticipated, it is worth working with equation (13.2.2.6) in the polar parameterization,

$$\mathrm{d}s^2 = \mathrm{d}\chi^2 + [2\sin(\chi/2)]^2\mathrm{d}\mathbf{u}^2. \qquad (13.2.3.17)$$

For constant χ, we have the topology of the spherical surface. It is then natural to plot the function as a stereographic projection.

In the case of the cross-rotation function, *i.e.*, when the atomic distribution of the search model is known, we may use equation (13.2.2.9) instead of equation (13.2.2.6) to define the sampling. For this, we choose σ as the smallest root-mean-square shift in atomic coordinates that produces an appreciable change in the value of the cross-rotation function. An acceptable sampling set of $\mathcal{R}$ should then satisfy the following conditions:

(1) For every point in the rotation domain, there is at least one sampling point at a distance less than σ.
(2) The number of sampling points is as small as possible.

13.2.3.3. Strategies

Since the first attempts to detect subunits within the crystal asymmetric unit, a great deal of experience has been gained and the results collected in several treatises on molecular replacement (Rossmann, 1972; Machin, 1985; Dodson *et al.*, 1992; Carter & Sweet, 1997).

The success of the rotation function relies in part on the choice of the domain of integration, *i.e.*, inner and outer radii of the spherical domain, and limits on data resolution. The outer radius is chosen so as to maximize the ratio between the number of intramolecular and intermolecular vectors. A typical value of this radius is 50–75% of the subunit's diameter for spherical molecules. The choice of the inner radius is less crucial, provided that the origin peaks are subtracted from the Patterson functions; it is often set to zero.

The resolution range of the selected diffraction data depends on whether we are computing self- or cross-rotation functions. For self rotations, data up to the highest available resolution may be used. For cross rotations, the high-resolution limit is dictated by the degree of similarity between the probe and the actual molecules that constitute the crystal. The low-resolution limit is usually chosen so as to skip the solvent contribution.

Another parameter of interest concerns the angular resolution, which is rather specific to the fast rotation function. Although it was derived by expanding the interference function to follow the original reciprocal-space formulation as closely as possible, it may also be obtained by expanding the Patterson functions in spherical harmonics and then performing the integration (Crowther, 1972). In this way, the concept of angular resolution is more evident. In fact, the spherical harmonics play the same role in the angular domain as the imaginary exponentials do in Cartesian coordinate space. The analogue of the Miller indexes here is the label ℓ, directly related to the angular resolution. The term with $\ell = 0$ is invariant under rotations and is thus eliminated from the summation in equation (13.2.3.14). It represents the contribution of the best radial functions, in a least-squares sense, that approximate the Patterson functions within Ω. The origin peak is thus properly removed by omitting the $\ell = 0$ term, as well as a substantial component of the original function. Also, by omitting low ℓ's, the angular resolution of peaks is enhanced. These eliminations may also be done in other formulations, but never as efficiently.

The parameters discussed above are not all independent. The relationship between them stems from the behaviour of the spherical Bessel functions for small values of their argument (Watson, 1958),

$$j_\ell(2\pi hb) \simeq (2\pi hb)^\ell/(2\ell+1)!!$$

[see equation (13.2.3.8)]. So, when omitting low ℓ terms, we are also weighting down low-resolution data.

13.2.3.4. Symmetry properties of the rotation function

The overlap integral that defines $\mathcal{R}$ may be calculated by rotating P_t instead of P_s, but with the inverse rotation,

$$\begin{aligned}\mathcal{R}(\mathbf{R}) &= (1/v)\int_{\Omega} P_t(\mathbf{r})P_s(\mathbf{R}^{-1}\mathbf{r})\;\mathrm{d}^3\mathbf{r}\\ &= (1/v)\int_{\Omega} P_t(\mathbf{R}\mathbf{r})P_s(\mathbf{r})\;\mathrm{d}^3\mathbf{r}. \qquad (13.2.3.18)\end{aligned}$$

This property enables the analysis of the consequence of the symmetries of the Patterson functions upon the rotation function (Tollin *et al.*, 1966; Moss, 1985). For example, when the target and search functions are the same, a trivial symmetry of the self-rotation results: its values at $\mathbf{R}$ and $\mathbf{R}^{-1}$ are the same or, in Euler angles, at (α, β, γ) and $(180-\gamma, \beta, 180-\alpha)$. More generally, if P_t is invariant under the rotation $\mathbf{T}$, *i.e.*, $P_t(\mathbf{T}^{-1}\mathbf{r}) = P_t(\mathbf{r})$ [and similarly for the search function, $P_s(\mathbf{S}^{-1}\mathbf{r}) = P_s(\mathbf{r})$], then

$$\begin{aligned}\mathcal{R}(\mathbf{R}) &= (1/v)\int_{\Omega} P_t(\mathbf{r})P_s(\mathbf{R}^{-1}\mathbf{r})\;\mathrm{d}^3\mathbf{r}\\ &= (1/v)\int_{\Omega} P_t(\mathbf{T}^{-1}\mathbf{r})P_s(\mathbf{S}^{-1}\mathbf{R}^{-1}\mathbf{r})\;\mathrm{d}^3\mathbf{r}\\ &= (1/v)\int_{\Omega} P_t(\mathbf{r})P_s(\mathbf{S}^{-1}\mathbf{R}^{-1}\mathbf{T}\;\mathbf{r})\;\mathrm{d}^3\mathbf{r} = \mathcal{R}(\mathbf{T}^{-1}\mathbf{R}\mathbf{S}).\end{aligned} \qquad (13.2.3.19)$$

However, the actual symmetry displayed by $\mathcal{R}$ depends on the parameterization of the rotations and on the orientation of the orthonormal axes with respect to the crystal ones. Within the Euler parameterization, if the target function has an n-fold rotation axis parallel to the orthonormal z axis, then, according to equation (13.2.2.3), $\mathcal{R}$ will have a periodicity of $360/n$ along α. Similarly, a rotation axis of order n along $\mathbf{z}$ of P_s gives rise to a periodicity of $360/n$ along γ. Therefore, the amount of calculation is reduced by choosing $\mathbf{z}$ along the Patterson functions' highest rotational symmetry axes [see equation (13.2.3.11)].

13.2.4. The locked rotation function

The rotational NCS, determined with the help of the self-rotation function, may be used to enhance the signal-to-noise ratio of cross-rotation functions (Rossmann *et al.*, 1972; Tong & Rossmann, 1990). If $\{\mathbf{S}_n, n = 1, \ldots, N\}$ denotes the set of NCS rotations, including the identity, and $\mathbf{R}$ is a correct orientation of the cross rotation, then $\mathbf{S}_n\mathbf{R}$ must also correspond to a correct orientation. Here we are assuming that the rotational NCS forms a group. Otherwise, either $\mathbf{S}_n\mathbf{R}$ or $\mathbf{S}_n^{-1}\mathbf{R}$, but not both, corresponds to another correct orientation. Therefore, a function may be defined, the locked cross rotation, whose values are the average of the cross-rotation values at orientations related by the NCS:

$$\mathcal{R}_{LC}(\mathbf{R}) = \sum_{n=1}^{N}\mathcal{R}(\mathbf{S}_n\mathbf{R})/N. \qquad (13.2.4.1)$$

By redefining the target function, it can be computed as an ordinary cross rotation. Indeed, $\mathcal{R}_{LC}$ may be written in a form similar to equation (13.2.3.3),

$$\begin{aligned}\mathcal{R}_{LC}(\mathbf{R}) &= \sum_{n=1}^{N}(1/v)\int_{\Omega} P_t(\mathbf{r})P_s(\mathbf{R}^{-1}\mathbf{S}_n^{-1}\mathbf{r})\ \mathrm{d}^3\mathbf{r}/N \\ &= (1/v)\int_{\Omega}\left[\sum_{n=1}^{N}P_t(\mathbf{S}_n\mathbf{r})/N\right]P_s(\mathbf{R}^{-1}\mathbf{r})\ \mathrm{d}^3\mathbf{r}, \qquad (13.2.4.2)\end{aligned}$$

with the target Patterson function substituted by the average over the NCS of the rotated target functions. The computation of equation (13.2.4.2) is particularly simple in the case of the fast rotation function. The substitution

$$e^{(t)}_{\ell,m,n} \rightarrow \sum_{m'=-\ell}^{\ell}\left[\sum_{n=1}^{N}\mathcal{D}^{\ell}_{m,m'}(\mathbf{S}_n)/N\right]e^{(t)}_{\ell,m',n}, \qquad (13.2.4.3)$$

where we replaced the sum over $\mathbf{S}_n^{-1}$ by a sum over $\mathbf{S}_n$, because of the rearrangement theorem of group theory, gives the required target coefficients.

The same ideas may be applied to the self-rotation function. Here the NCS is assumed beforehand, with elements $\{\mathbf{I}_n, n = 1, \ldots, N\}$ in a given reference orientation. They are related to the actual NCS elements by

$$\mathbf{S}_n = \mathbf{R}_n\mathbf{I}_n\mathbf{R}_n^{-1} \qquad (13.2.4.4)$$

[see equation (13.2.2.2)]. Since each $\mathbf{S}_n$ should correspond to a local maximum of the self rotation, the function

$$\mathcal{R}_{LS}(\mathbf{R}) = \sum_{n=1}^{N}\mathcal{R}(\mathbf{R}\mathbf{I}_n\mathbf{R}^{-1})/N \qquad (13.2.4.5)$$

should also display a maximum for each $\mathbf{R}_n$, but with a noise level reduced by $(N)^{1/2}$. Equation (13.2.4.5) defines the locked self-rotation function. Its computation by fast rotation techniques is less straightforward than in the locked cross-rotation case.

13.2.5. Other rotation functions

The rotation function was hitherto described in terms of self- and cross-Patterson vectors. This is perhaps inevitable in the self-rotation case, but the problem of determining the absolute orientation of the subunits when a model structure is available may be formulated in a different way. We may try to compare directly the observed and calculated intensities or structure factors by using any criterion analogous to those employed in refinement procedures, *e.g.*, the crystallographic R factor or correlation coefficients.

When the space-group symmetry is explicitly exhibited, the structure factor corresponding to a crystal with M independent molecules in the unit cell takes the form

$$F(\mathbf{h}) = \sum_{m=1}^{M}\sum_{g=1}^{G} f_m(\mathbf{h}\mathbf{M}_g)\exp[2\pi i\mathbf{h}(\mathbf{M}_g\mathbf{r}_m + \mathbf{t}_g)], \qquad (13.2.5.1)$$

where $\mathbf{M}_g$ and $\mathbf{t}_g$ denote, respectively, the transformation matrix and the translation associated with the gth symmetry operation of the crystal space group. The corresponding intensity is

$$\begin{aligned}I(\mathbf{h}) = \sum_{m,m'=1}^{M}\ \sum_{g,g'=1}^{G}\overline{f_m(\mathbf{h}\mathbf{M}_g)}f_{m'}(\mathbf{h}\mathbf{M}_{g'})& \\ \times\exp[2\pi i\mathbf{h}(\mathbf{M}_g\mathbf{r}_m + \mathbf{t}_g - \mathbf{M}_{g'}\mathbf{r}_{m'} - \mathbf{t}_{g'})].& \qquad (13.2.5.2)\end{aligned}$$

For criteria based on amplitudes, the calculated structure factor will contain only the contribution of the rotated model,

$$F^{\text{calc}}_{\mathbf{h}}(\mathbf{R}) = f_m(\mathbf{h}\mathbf{R}), \qquad (13.2.5.3)$$

i.e., the Fourier transform of a single molecule in the crystal cell, assuming $P1$ symmetry. For criteria based on intensities, some symmetry information may be introduced,

$$I^{\text{calc}}_{\mathbf{h}}(\mathbf{R}) = \sum_{g=1}^{G}|\,f_m(\mathbf{h}\mathbf{M}_g\mathbf{R})|^2. \qquad (13.2.5.4)$$

A criterion often considered is the correlation coefficient on intensities,

$$\begin{aligned}\mathcal{R}_{\mathcal{D}}(\mathbf{R}) &= \langle(I^{\text{obs}} - \langle I^{\text{obs}}\rangle)(I^{\text{calc}} - \langle I^{\text{calc}}\rangle)\rangle \\ &\quad\times[\langle(I^{\text{obs}} - \langle I^{\text{obs}}\rangle)^2\rangle\langle(I^{\text{calc}} - \langle I^{\text{calc}}\rangle)^2\rangle]^{1/2}, \qquad (13.2.5.5)\end{aligned}$$

where $\langle\ldots\rangle$ means 'average over reflections'. It may be calculated within reasonable computing time provided that

(1) the structure factors are computed by interpolation from the Fourier transform of the isolated molecule's electron density; and

(2) an efficient sampling set of $\mathcal{R}_{\mathcal{D}}$ is defined.

$\mathcal{R}_{\mathcal{D}}$ is referred to as the direct-rotation function (DeLano & Brünger, 1995). A major advantage of this formulation is that the information stemming from already-positioned subunits may be taken into account, just by adding their contribution to the calculated intensities.

13.2.6. Concluding remarks

Each formulation of the rotation function described above has its advantages and disadvantages. The direct-space formulation [equation (13.2.3.3)] offers the possibility of modifying the Patterson function or selecting the strongest peaks to be used in the overlap integral. Also, the domain of integration may have any shape, as in the reciprocal-space formulation [equation (13.2.3.5)]. However, in both formulations, the numerical results can be somewhat imprecise because of the approximations introduced to save computing time.

When the domain of integration is spherical, as is usually the case, then the fast rotation function [equation (13.2.3.15)] is faster, more accurate and allows for angular-resolution enhancement. Moreover, since most of the computing time is spent in the calculation of the coefficients $e_{\ell,m,n}$, a library of these coefficients may be compiled to assess the models in a given molecular-replacement problem more rapidly.

The direct rotation [equation (13.2.5.5)] seems preferable to the cross-rotation function, as it includes all self-Patterson terms of the search model. However, when the order of the space group is high, the calculated intensity represents only a small fraction of the observed one, and the discriminative power of the function drops, as compared with the cross-rotation function.

Patterson searches now benefit from new supercomputers. The real problem of molecular replacement, split for convenience into rotation and translation searches, is beginning to be tackled by genuine six-dimensional searches where all orientations are tested. This may represent the end of cross-rotation and direct-rotation functions.

APPENDIX A13.2.1

Formulae for the derivation and computation of the fast rotation function

This appendix aims to present a complete set of formulae which allow the derivation and computation of the fast rotation function. They involve a particular convention for the definition of the irreducible representations of the rotation group suitable for crystallographic computations.

A13.2.1.1. Euler parameterization

By applying the group property of rotations [equation (13.2.2.2)], the Euler parameterization may be expressed as rotations around fixed axis (see Fig. 13.2.2.1*b*):

$$\begin{aligned}\mathbf{R}(\alpha,\beta,\gamma) &= \mathbf{R}(\gamma,\mathbf{p})\mathbf{R}(\beta,\mathbf{n})\mathbf{R}(\alpha,\mathbf{z})\\ &= \left[\mathbf{R}(\beta,\mathbf{n})\mathbf{R}(\gamma,\mathbf{z})\mathbf{R}(\beta,\mathbf{n})^{-1}\right]\mathbf{R}(\beta,\mathbf{n})\mathbf{R}(\alpha,\mathbf{z})\\ &= \mathbf{R}(\beta,\mathbf{n})\mathbf{R}(\gamma,\mathbf{z})\mathbf{R}(\alpha,\mathbf{z})\\ &= \left[\mathbf{R}(\alpha,\mathbf{z})\mathbf{R}(\beta,\mathbf{y})\mathbf{R}(\alpha,\mathbf{z})^{-1}\right]\mathbf{R}(\gamma,\mathbf{z})\mathbf{R}(\alpha,\mathbf{z})\\ &= \mathbf{R}(\alpha,\mathbf{z})\mathbf{R}(\beta,\mathbf{y})\mathbf{R}(\gamma,\mathbf{z}). \qquad \text{(A13.2.1.1)}\end{aligned}$$

A13.2.1.2. The $D^{\ell}_{m,m'}$ matrices

A linear representation of dimension n of the rotation group is a correspondence between rotations and matrices of order n. The matrices $D^{\ell}_{m,m'}(\mathbf{R})$, with $-\ell \le m, m' \le \ell$, are associated with the irreducible representation of dimension $2\ell+1$ $(0 \le \ell < \infty)$. They have the following properties (Brink & Satchler, 1968):

(1) Group multiplication:

$$D^{\ell}_{m,m'}(\mathbf{RR}') = \sum_{n=-\ell}^{\ell} D^{\ell}_{m,n}(\mathbf{R}) D^{\ell}_{n,m'}(\mathbf{R}'). \qquad \text{(A13.2.1.2)}$$

(2) Complex conjugation:

$$D^{\ell}_{m,m'}(\mathbf{R}^{-1}) = \overline{D^{\ell}_{m',m}(\mathbf{R})}. \qquad \text{(A13.2.1.3)}$$

(3) Euler parameterization:

$$D^{\ell}_{m,m'}(\alpha,\beta,\gamma) = d^{\ell}_{m,m'}(\beta)\exp[i(m\alpha + m'\gamma)]. \qquad \text{(A13.2.1.4)}$$

(4) Recurrence relation for the reduced matrices:

$$\begin{aligned} d^{\ell}_{m-1,m'}(\beta) &= \frac{2[m' - m\cos(\beta)]}{[(\ell-m+1)(\ell+m)]^{1/2}\sin(\beta)} d^{\ell}_{m,m'}(\beta)\\ &\quad - \left[\frac{(\ell-m)(\ell+m+1)}{(\ell-m+1)(\ell+m)}\right]^{1/2} d^{\ell}_{m+1,m'}(\beta). \qquad \text{(A13.2.1.5)}\end{aligned}$$

(5) Initial values (bottom row of d^{ℓ}):

$$d^{\ell}_{\ell,m}(\beta) = (-1)^{\ell-m}\left[\frac{(2\ell)!}{(\ell-m)!(\ell+m)!}\right]^{1/2} \sin(\beta/2)^{\ell-m}\cos(\beta/2)^{\ell+m}. \qquad \text{(A13.2.1.6)}$$

(6) Symmetry relations:

$$d^{\ell}_{-m,-m'}(\beta) = d^{\ell}_{m',m}(\beta) = (-1)^{m-m'} d^{\ell}_{m,m'}(\beta). \qquad \text{(A13.2.1.7)}$$

A13.2.1.3. Spherical harmonics

The $Y_{\ell,m}$'s, with $-\ell \le m \le \ell$ and $0 \le \ell < \infty$, constitute a complete set of functions of the unit vector $\mathbf{u}$, having the following properties (Brink & Satchler, 1968):

(1) Transformation under rotations:

$$\begin{aligned} Y_{\ell,m}(\mathbf{R}^{-1}\mathbf{u}) &= \sum_{m'=-\ell}^{\ell} D^{\ell}_{m,m'}(\mathbf{R}^{-1}) Y_{\ell,m'}(\mathbf{u})\\ &= \sum_{m'=-\ell}^{\ell} Y_{\ell,m'}(\mathbf{u})\overline{D^{\ell}_{m',m}(\mathbf{R})}. \qquad \text{(A13.2.1.8)}\end{aligned}$$

(2) Orthogonality condition:

$$\int \overline{Y_{\ell,m}(\mathbf{u})} Y_{\ell',m'}(\mathbf{u})\,\mathrm{d}^2\mathbf{u} = \delta_{\ell,\ell'}\delta_{m,m'}. \qquad \text{(A13.2.1.9)}$$

(3) Inversion:

$$Y_{\ell,m}(-\mathbf{u}) = (-1)^{\ell}\; Y_{\ell,m}(\mathbf{u}). \qquad \text{(A13.2.1.10)}$$

(4) Relation with rotation-matrix elements:

$$Y_{\ell,m}(\theta,\varphi) = i^{\ell}[(2\ell+1)/4\pi]^{1/2} D^{\ell}_{m,0}(\varphi,\theta,0), \qquad \text{(A13.2.1.11)}$$

where (θ, φ) are the polar coordinates of $\mathbf{u}$.

A13.2.1.4. Spherical Bessel functions

The j_{ℓ}'s, with $0 \le \ell < \infty$, constitute a complete set of functions having the following properties (Watson, 1958):

(1) Recurrence relation:

$$j_{\ell-1}(x) - (2\ell+1)[\,j_{\ell}(x)/x] + j_{\ell+1}(x) = 0. \qquad \text{(A13.2.1.12)}$$

(2) Initial values:

$$\begin{aligned} j_0(x) &= \sin(x)/x\\ j_1(x) &= [\sin(x) - x\ \cos(x)]/x^2. \qquad \text{(A13.2.1.13)}\end{aligned}$$

(3) Integral of a product of spherical Bessel functions:

$$\begin{aligned} U^{\ell}(p,q) &= \int_0^1 j_{\ell}(px) j_{\ell}(qx) x^2\,\mathrm{d}x\\ &= \begin{cases} [\,j_{\ell}(p) j_{\ell-1}(q) q - j_{\ell}(q) j_{\ell-1}(p) p]/(p^2 - q^2) & \text{if } p \ne q\\ \tfrac{1}{2}[\,j_{\ell}(p)^2 - j_{\ell-1}(p) j_{\ell+1}(p)] & \text{if } p = q \end{cases}\\ &= (2\ell+3)[\,j_{\ell+1}(p) j_{\ell+1}(q)]/pq + U^{\ell+2}(p,q)\\ &= \sum_{n=1}^{\infty}[2(\ell+2n) - 1][\,j_{\ell+2n-1}(p) j_{\ell+2n-1}(q)]/pq. \qquad \text{(A13.2.1.14)}\end{aligned}$$

A13.2.1.5. Expansion of $\exp(2\pi i\mathbf{sr})$

This is also called the plane-wave expansion or Laplace's expansion (Landau & Lifschitz, 1972):

$$\exp(2\pi i\mathbf{sr}) = 4\pi\sum_{\ell=0}^{\infty} i^{\ell} \sum_{m=-\ell}^{\ell} j_{\ell}(2\pi s r)\overline{Y_{\ell,m}(\hat{\mathbf{s}})} Y_{\ell,m}(\hat{\mathbf{r}}). \qquad \text{(A13.2.1.15)}$$

A13.2.1.6. Expansion of the interference function

$$\begin{aligned}
\chi_b(\mathbf{h}-\mathbf{k}\mathbf{R}^{-1}) &= (3/4\pi b^3)\int_0^b\int_0^\pi\int_0^{2\pi} \exp[2\pi i(\mathbf{h}-\mathbf{k}\mathbf{R}^{-1})\mathbf{r}]r^2\sin(\theta)\,\mathrm{d}r\,\mathrm{d}\theta\,\mathrm{d}\varphi \\
&= \sum_{\ell,\ell'=0}^{\infty}\sum_{m=-\ell}^{\ell}\sum_{m'=-\ell'}^{\ell'} i^{\ell-\ell'}\overline{Y_{\ell,m}(\hat{\mathbf{h}})}Y_{\ell',m'}(\hat{\mathbf{k}}) \\
&\quad\times(12\pi/b^3)\int_0^b j_\ell(2\pi hr)\,j_{\ell'}(2\pi kr)r^2\,\mathrm{d}r \\
&\quad\times\int_0^\pi\int_0^{2\pi} Y_{\ell,m}(\hat{\mathbf{r}})\overline{Y_{\ell',m'}(\mathbf{R}^{-1}\hat{\mathbf{r}})}\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\varphi \\
&= \sum_{\ell=0}^{\infty}\sum_{m=-\ell}^{\ell}\sum_{\ell'=0}^{\infty}\sum_{m'=-\ell'}^{\ell'} i^{\ell-\ell'}\overline{Y_{\ell,m}(\hat{\mathbf{h}})}Y_{\ell',m'}(\hat{\mathbf{k}}) \\
&\quad\times(12\pi/b^3)\int_0^b j_\ell(2\pi hr)\,j_{\ell'}(2\pi kr)r^2\,\mathrm{d}r \\
&\quad\times\sum_{m''=-\ell'}^{\ell'}\int_0^\pi\int_0^{2\pi} Y_{\ell,m}(\hat{\mathbf{r}})\overline{Y_{\ell',m''}(\hat{\mathbf{r}})}\sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\varphi\; D^{\ell'}_{m'',m'}(\mathbf{R}) \\
&= \sum_{\ell=0}^{\infty}\sum_{m'=-\ell}^{\ell}\overline{Y_{\ell,m}(\hat{\mathbf{h}})}Y_{\ell,m'}(\hat{\mathbf{k}})12\pi U^\ell(2\pi hb,2\pi kb)D^\ell_{m,m'}(\mathbf{R}).
\end{aligned}\tag{A13.2.1.16}$$

References

Brink, D. M. & Satchler, G. R. (1968). *Angular Momentum*, 2nd ed. Oxford University Press.

Burdina, V. I. (1971). *Symmetry of rotation function. Sov. Phys. Crystallogr.* **15**, 545–550.

Carter, C. W. & Sweet, R. M. (1997). *Molecular replacement. Methods Enzymol.* **276**, 558–619.

Crowther, R. A. (1972). In *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon and Breach.

DeLano, W. L. & Brünger, A. T. (1995). *The direct rotation function: rotational Patterson correlation search applied to molecular replacement. Acta Cryst.* D**51**, 740–748.

Dodson, E. J. (1985). In *Proceedings of the Daresbury Study Weekend. Molecular Replacement*, edited by P. A. Machin, pp. 33–45. Warrington: Daresbury Laboratory.

Dodson, E. J., Gover, S. & Wolf, W. (1992). Editors. *Proceedings of the Daresbury Study Weekend. Molecular Replacement*. Warrington: Daresbury Laboratory.

Landau, L. D. & Lifschitz, E. M. (1972). *Théorie Quantique Relativiste*, pp. 109–196. Moscow: Editions MIR.

Lattman, E. E. (1972). *Optimal sampling of the rotation function. Acta Cryst.* B**28**, 1065–1068.

Machin, P. A. (1985). Editor. *Proceedings of the Daresbury study weekend. Molecular replacement*. Warrington: Daresbury Laboratory.

Moss, D. S. (1985). *The symmetry of the rotation function. Acta Cryst.* A**41**, 470–475.

Navaza, J. (1993). *On the computation of the fast rotation function. Acta Cryst.* D**49**, 588–591.

Nordman, C. E. (1966). *Vector space search and refinement procedures. Trans. Am. Crystallogr. Assoc.* **2**, 29–38.

Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method*. New York: Gordon and Breach.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Rossmann, M. G., Ford, G. C., Watson, H. C. & Banaszak, L. J. (1972). *Molecular symmetry of glyceraldehyde-3-phosphate dehydrogenase. J. Mol. Biol.* **64**, 237–245.

Steigemann, W. (1974). *Die Entwicklung und Anwendung von Rechenverfahren und Regenprogrammen zur Strukturanalyse von Proteinen am beispiel des Trypsin–Trypsininhibitor Komplexes, des Freien und der* L*-Asparaginase*. PhD thesis, Technische Universität, Munich, Germany.

Tollin, P., Main, P. & Rossmann, M. G. (1966). *The symmetry of the rotation function. Acta Cryst.* **20**, 404–417.

Tong, L. & Rossmann, M. G. (1990). *The locked rotation function. Acta Cryst.* A**46**, 783–792.

Watson, G. N. (1958). *A Treatise on the Theory of Bessel Functions*, 2nd ed. Cambridge University Press.

International Tables for Crystallography (2012). Vol. F, Chapter 13.3, pp. 347–351.

Chapter 13.3. Translation functions

L. Tong

13.3.1. Introduction

A structure determination by the molecular-replacement method traditionally proceeds in two steps (Rossmann, 1972, 1990, 2001). The first step involves the determination of the orientation of the search model in the crystal unit cell by the rotation functions (see Chapter 13.2). Once the orientation of the search model is known, translation functions are employed in the second step to determine the location of the model in the crystal unit cell. This essentially reduces a six-dimensional problem (three rotational and three translational degrees of freedom) to two three-dimensional problems, which are computationally more manageable. With the speed of modern computers, a strict division between the rotational and the translational components of a molecular-replacement structure solution may no longer be necessary (see Section 13.3.7).

Translation functions are normally formulated to achieve minimum or maximum values when the search model is at its correct position in the crystal unit cell. As with the rotation problem, the translation problem is frequently solved as a search. The positional parameters of the model are varied in the unit cell, generally on a grid. Translation functions are evaluated at these search grid points in order to identify those that minimize or maximize the functions.

Most translation functions involve a comparison between the observed structure-factor amplitudes (or squared amplitudes) and those calculated based on the search model. The *R* factor and the correlation coefficient can be used as indicators for translation searches (Section 13.3.2). The correlation between the observed Patterson map and that based on the search model is the foundation of another translation function (Section 13.3.3). If phase information is available from other sources, the correlation between the electron-density maps is the basis for phased translation functions (Section 13.3.4). Proper packing of the search model in the crystal unit cell is an essential component of a solution to the translation problem (Section 13.3.5). The power of the translation functions can be enhanced in the presence of noncrystallographic symmetry (Section 13.3.8).

Besides *R*-factor, correlation-coefficient and Patterson-correlation functions, target functions based on the principle of maximum likelihood have been found to be powerful indicators for molecular-replacement calculations (Read, 2001; McCoy *et al.*, 2005, 2007), and have been successful in helping determine structures where traditional molecular-replacement target functions have failed.

13.3.2. *R*-factor and correlation-coefficient translation functions

The crystallographic *R* factor is often used as an indicator in translation searches. It is a measure of the percentage difference between the observed (F_h^o) and the calculated (F_h^c) structure-factor amplitudes,

$$R_F = 100 \times \sum_h \left|F_h^o - k_F F_h^c\right| \Big/ \sum_h F_h^o. \qquad (13.3.2.1)$$

A similar *R* factor can be defined based on the square of the structure-factor amplitudes, *i.e.* an *R* factor based on intensity,

$$R_I = 100 \times \sum_h \left|I_h^o - k_I I_h^c\right| \Big/ 2\sum_h I_h^o. \qquad (13.3.2.2)$$

A factor of 2 is introduced in equation (13.3.2.2) to make R_I values fall in the same range as R_F. In equations (13.3.2.1) and (13.3.2.2) k_F and k_I are scale factors that bring the observed and the calculated structure factors to the same level. If these scale factors are calculated in shells of equal reciprocal volume, they can compensate for differences in the displacement factors between the observed and the calculated structure-factor amplitudes.

In addition to the *R* factor, the correlation coefficient between the observed and the calculated structure factors is also used in translation searches. Like the *R* factors, correlation coefficients can be defined based on the amplitude or the intensity of the reflections, and that based on the amplitude is shown below.

$$\begin{aligned} cc_F &= \frac{\sum_h \left(F_h^o - \langle F_h^o\rangle\right)\left(F_h^c - \langle F_h^c\rangle\right)}{\left[\sum_h \left(F_h^o - \langle F_h^o\rangle\right)^2 \left(F_h^c - \langle F_h^c\rangle\right)^2\right]^{1/2}} \\ &= \frac{\sum_h F_h^o F_h^c - \sum_h F_h^o \sum_h F_h^c / N}{\left\{\left[\sum_h \left(F_h^o\right)^2 - \left(\sum_h F_h^o\right)^2/N\right]\left[\sum_h \left(F_h^c\right)^2 - \left(\sum_h F_h^c\right)^2/N\right]\right\}^{1/2}}. \end{aligned} \qquad (13.3.2.3)$$

In (13.3.2.3), N is the number of reflections that are used in the calculation and $\langle F_h\rangle$ denotes the average structure-factor amplitude over the reflections. Unlike the *R* factors, the correlation coefficients do not depend on the overall scale factor between the observed and the calculated structure factors. However, they can be affected by large differences in the overall displacement factors between the observed and the calculated structure factors. The correlation coefficient based on intensity can be calculated by the fast Fourier transform (FFT) technique for translation-function calculations (Navaza & Vernoslova, 1995).

In order to evaluate *R* factors and correlation coefficients for a translation search, structure factors need to be calculated for the search model, with a given orientation, at different positions in the unit cell. For this special case, where only the positional parameters of the search model are varied, the calculation of the structure factors can be simplified (Nixon & North, 1976; Rae, 1977). The structure-factor equation can be written as a double summation – first over the atoms in one asymmetric unit of the unit cell and then over all the asymmetric units,

$$\bar{F}_h^c = \sum_n \sum_j f_j \exp\left\{2\pi i h\left(\left[T_n\right]x_j + t_n\right)\right\}, \qquad (13.3.2.4)$$

where j goes over all the atoms in the asymmetric unit and n goes over all the crystallographic asymmetric units. The nth crystallographic symmetry operator is given by

$$x_n = \left[T_n\right]x + t_n, \qquad (13.3.2.5)$$

where $\left[T_n\right]$ is the rotational component and t_n is the translational component of the symmetry operator.

For simplicity, first consider the case where there is only one molecule in the asymmetric unit. In the translation search, the model will be placed at different positions in the unit cell,

Copyright © 2012 International Union of Crystallography

$$x_j = x_j^0 + v_0, \tag{13.3.2.6}$$

where v_0 is a translation vector which is applied to move the model from its starting position (x_j^0). Substituting x_j in (13.3.2.6) into (13.3.2.4) gives

$$\bar{F}_h^c = \sum_n \bar{f}_{h,n} \exp\big(2\pi i h\big[T_n\big]v_0\big), \tag{13.3.2.7}$$

where $\bar{f}_{h,n}$ is the structure factor calculated based only on the nth symmetry-related molecule,

$$\bar{f}_{h,n} = \sum_j f_j \exp\big\{2\pi i h\big(\big[T_n\big]x_j^0 + t_n\big)\big\}. \tag{13.3.2.8}$$

It can be calculated by placing the search model at the starting position in a $P1$ unit cell having the same cell dimensions as the crystal unit cell. The structure factors calculated for this $P1$ cell are related to $\bar{f}_{h,n}$ by

$$\bar{f}_{h,n} = \exp(2\pi i h t_n)\bar{f}_{h[T_n],\,1}. \tag{13.3.2.9}$$

Therefore, the summation over the atoms in the structure-factor calculation, a rather time-consuming process, needs to be performed only once, for the search model at the starting position. Subsequent structure-factor calculations after translation of the model are no longer dependent on the number of atoms present in the unit cell [equation (13.3.2.7)]. The starting position is usually chosen such that the centre of the search model is at (0, 0, 0). Then the vector that is determined from the translation searches will define the centre of the model in the unit cell.

Equation (13.3.2.7) can be generalized to allow for the presence of other molecules that are to remain stationary during the translation search:

$$\bar{F}_h^c = \bar{A}_h + \sum_n \bar{f}_{h,n} \exp(2\pi i h[T_n]v_0), \tag{13.3.2.10}$$

where $\bar{A}_h$ is the contribution from the stationary molecules. This formulation is useful if there is more than one molecule in the asymmetric unit. The position of one of the molecules can be determined first, and the model is then included as a stationary molecule for the position search of the next molecule. Incorporation of the contribution from stationary molecules can also improve the quality of the translation function (Zhang & Matthews, 1994).

Equation (13.3.2.7) can also be generalized to allow for the presence of two (or more) search models that are to move independently of each other during the translation search. However, this will generally lead to a six- (or more) dimensional problem and is expensive in computation time. With improvements in computer technology (especially parallel processing), it may be feasible to carry out such searches in special cases. However, this aspect of translation functions will not be discussed further here.

13.3.3. Patterson-correlation translation function

Besides the R factor and the correlation coefficient, another commonly used translation search indicator is based on the correlation between the observed and the calculated Patterson maps (Crowther & Blow, 1967; Tong, 1993; Navaza & Vernoslova, 1995). Rotation functions are based on the overlap of only a subset of the interatomic vectors in the Patterson map, *i.e.* only those near the origin of the unit cell, which generally contain the self vectors within each molecule. The correct orientation and position of a search model in the crystal unit cell should lead to the maximal overlap of both the self and the cross vectors, *i.e.* maximal overlap between the observed (P^o) and the calculated (P^c) Patterson maps throughout the entire unit cell (Tong, 1993),

$$\begin{aligned} PC &= \int_\Omega P^o(u)P^c(u)\,\mathrm{d}u \\ &= \sum_h (F_h^o)^2 |\bar{F}_h^c|^2. \end{aligned} \tag{13.3.3.1}$$

The calculated structure factor is a function of the translation vector v_0 [equation (13.3.2.10)]. Combining equations (13.3.2.10) and (13.3.3.1) gives

$$\begin{aligned} PCTF(v_0) = {} & \sum_h (F_h^o)^2 \big|A_h\big|^2 + \sum_h \sum_n (F_h^o)^2 \big|\bar{f}_{h,n}\big|^2 \\ & + 2\sum_h \sum_n (F_h^o)^2 \bar{A}_h \bar{f}_{h,n}^* \exp(-2\pi i h[T_n]v_0) \\ & + \sum_h \sum_n \sum_{m\neq n} (F_h^o)^2 \bar{f}_{h,n} \bar{f}_{h,m}^* \exp\big\{-2\pi i h\big(\big[T_m\big] - \big[T_n\big]\big)v_0\big\}. \end{aligned} \tag{13.3.3.2}$$

The first two terms in equation (13.3.3.2) contribute a constant to the correlation and are generally ignored in this calculation (but see Section 13.3.8). The Patterson-correlation translation function is therefore a Fourier transform, with $h[T_n]$ (from the third term) and $h([T_m] - [T_n])$ (from the fourth term) as the indices. This function can be evaluated quickly by the FFT technique. The term $h([T_m] - [T_n])$ often leads to the doubling of the original reflection indices. For example, $h([T_1] - [T_2])$ will give rise to ($2h$, 0, $2l$) for monoclinic space groups (b-unique setting). Therefore, the Patterson-correlation Fourier transform should normally be sampled with a grid size about 1/6 of the maximum resolution of the reflection data used in the calculation.

A disadvantage of the Patterson-correlation translation function as formulated in equation (13.3.3.2) is that the results from the calculation are on an arbitrary scale. This makes it difficult to compare results from different calculations. The R factor or the correlation coefficient can be calculated for the top peaks in the Patterson-correlation translation function [equation (13.3.3.2)] to place the results on an 'absolute' scale. Alternatively, the correlation, as defined by equation (13.3.3.1), can be normalized to become a Patterson-correlation coefficient (Harada *et al.*, 1981),

$$PC = \frac{\sum_h \big(F_h^o\big)^2 \big|\bar{F}_h^c\big|^2}{\Big[\sum_h \big(F_h^o\big)^4 \sum_h \big|\bar{F}_h^c\big|^4\Big]^{1/2}}. \tag{13.3.3.3}$$

This correlation coefficient is equivalent to that defined by equation (13.3.2.3) for reflection intensities (Fujinaga & Read, 1987). The evaluation of the $(\sum_h |\bar{F}_h^c|^4)^{1/2}$ term in the denominator of equation (13.3.3.3) is, however, time consuming, since the FFT technique cannot be applied directly. This term can be approximated by an expression which also measures the packing of the search molecules in the crystal (see Section 13.3.5).

The intramolecular vectors can be removed (Crowther & Blow, 1967) from the Patterson maps by subtracting, after appropriate scaling, the structure factors calculated from individual molecules [equation (13.3.2.8)].

The translation functions, as defined above, are based on the structure-factor amplitudes. Normalized structure factors (E's) may provide better results under certain circumstances, since they increase the weight of the high-resolution reflections in the translation function (Tollin, 1966; Harada *et al.*, 1981; Tickle, 1985). Since the Patterson-correlation translation function is also based on reflection intensities, the 'large term' approach can be used to accelerate the calculation (Tollin & Rossmann, 1966).

13.3.4. Phased translation function

If an atomic model needs to be placed in an electron-density map that has been obtained through other methods (*e.g.* the multiple-isomorphous-replacement method or partial model phases), the phased translation function can be used (Colman *et al.*, 1976; Read & Schierbeek, 1988; Bentley & Houdusse, 1992; Tong, 1993; Zhang & Matthews, 1994). It is essentially based on the correlation between observed and calculated electron-density values throughout the unit cell:

$$PTF(v_0) = \sum_h \sum_n \bar{F}^o_h \bar{f}^*_{h,n} \exp\big(-2\pi i h\big[T_n\big]v_0\big). \qquad (13.3.4.1)$$

Therefore, the phased translation function can also be evaluated by the FFT technique. As with the Patterson-correlation translation function, the phased translation function can be placed on an 'absolute' scale by introducing appropriate normalizing factors or by converting the results to R factors or correlation coefficients. It should be noted that the prior phase could be in the wrong hand, so both $\bar{F}^o_h$ and $(\bar{F}^o_h)^*$ may need to be tried in the phased translation function, especially if the phase information is derived from isomorphous replacement.

The stationary molecules contribute a constant to the phased translation function and are not shown in equation (13.3.4.1). However, the phase information from the stationary molecules can be applied to the observed structure-factor amplitudes, and the phased translation function, rather than the Patterson-correlation translation function, can be used in the search for additional molecules (Read & Schierbeek, 1988; Driessen *et al.*, 1991; Bentley & Houdusse, 1992). This could prove especially useful in locating the last few molecules in cases where there are several molecules in the asymmetric unit. Modified phased translation functions have been designed to locate fragments in the electron-density map (Cowtan, 1998).

13.3.5. Packing check in translation functions

A correct molecular-replacement solution should lead to the placement of the search model at the correct orientation and position in the crystal unit cell. For this solution, there should be no or minimal steric clashes among the crystallographically related and noncrystallographically related molecules in the unit cell (unless there are large conformational changes for the model in the crystal unit cell). Therefore, proper packing of the search model in the crystal unit cell is an important component of the molecular-replacement structure solution.

The packing of the search model in the unit cell can be estimated by determining the electron-density overlap among the molecules. This overlap can be calculated numerically, given the molecular envelope (Hendrickson & Ward, 1976). It can also be estimated by an analytical function (Harada *et al.*, 1981),

$$O(v_0) = \sum_h \big|\bar{F}^c_h(v_0)\big|^2 \Big/ N \sum_h \big|\bar{f}_{h,1}\big|^2, \qquad (13.3.5.1)$$

where N is the number of crystallographic symmetry operators. This overlap function assumes a value of 1 when there is no overlap among the molecules, and higher values when there is overlap. This function has been used to replace the $(\sum_h |\bar{F}^c_h|^4)^{1/2}$ term in the denominator of equation (13.3.3.3) (Harada *et al.*, 1981). Consequently, those positions that lead to steric clashes among the molecules will be down-weighted, thereby increasing the signal for the correct solution.

The overlap functions provide an overall estimate for the packing of the search model in the unit cell. A more detailed packing analysis can be based on the checking of atomic contacts (Tong, 1993). For example, the number of $C_\alpha \cdots C_\alpha$ contacts below a pre-specified distance cutoff (normally between 2 to 3 Å) in a protein crystal can be determined. Too many such contacts would indicate significant overlap of the molecules. For nucleic acid structures, a set of representative atoms (for example, P, C_1') can be selected from each nucleotide for this packing analysis.

13.3.6. The unique region of a translation function (the Cheshire group)

The region of the unit cell that should be covered during a translation search does not generally correspond to the asymmetric unit of the space group. Since the search model has a defined orientation, it can only reside in one of the asymmetric units in the unit cell. Lacking knowledge as to which asymmetric unit the model occupies, the entire unit cell would need to be searched. However, most space groups possess alternative origins, which means the position of a molecule in the unit cell can only be determined to within certain sets of translations. For example, in space group $P2_12_12_1$, there are eight alternative origins at $(0,0,0)$, $(\frac{1}{2},0,0)$, $(0,\frac{1}{2},0)$, $(0,0,\frac{1}{2})$, $(\frac{1}{2},\frac{1}{2},0)$, $(\frac{1}{2},0,\frac{1}{2})$, $(0,\frac{1}{2},\frac{1}{2})$ and $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$. This implies that the region that should be searched to locate a molecule need only be $\frac{1}{8}$ of the volume of the unit cell [for example, $(\frac{1}{2},\frac{1}{2},\frac{1}{2})$]. In addition, for polar space groups, the position of the molecule along the polar axis is arbitrary. The symmetry, as defined by these unique regions, is also known as the Cheshire group (Hirshfeld, 1968), and has been defined for all the 230 space groups.

Once the first molecule is positioned, the origin of the unit cell is fixed as well. The search for subsequent molecules will need to cover the entire unit cell, unless the crystal lattice is centred.

13.3.7. Combined molecular replacement

The traditional division of the molecular-replacement problem into two steps is partly due to limited computer power. Such a division has placed more pressure on the rotation function, since generally only a few rotation angles are examined by translation functions. The correct orientation, therefore, needs to be among the top few peaks either directly in the rotation functions or after Patterson-correlation refinement (Brünger, 1990).

With modern computers, it is no longer necessary to maintain the strict division between the rotational and the translational components. Even though a full six-dimensional search is still generally impractical, a limited six-dimensional search can certainly be performed (Navaza, 1994; Tong, 1996*a*; Sheriff *et al.*, 1999; Jogl *et al.*, 2001). Target functions that can be evaluated quickly with the FFT technique are preferred for this purpose. Otherwise, the exploration of rotational degrees of freedom (Fujinaga & Read, 1987) or the reflection data that are used in the calculation (Rabinovich & Shakked, 1984) will be limited.

An automated molecular-replacement protocol has been implemented which automatically examines the top peaks in the rotation function by translation functions (Navaza, 1994). This protocol has proven to be remarkably powerful. It assumes that the correct rotation solution is near the top peaks in the rotation function. A more general assumption is that the correct rotation for the search model should produce high values in the rotation function, even though they may not be near peaks in the rotation function. An error of 6° between the correct rotation angles and the peak in the rotation function, which often occurs, can make it impossible to obtain the correct translation-function solution (Fujinaga & Read, 1987).

The combined-molecular-replacement protocol (Tong, 1996*a*; Jogl *et al.*, 2001) therefore consists of examining all the grid points in the rotation function with heights greater than a defined cutoff value using the Patterson-correlation translation function. The top peaks (usually 20 to 50) in each translation function are all examined as possible solutions. The results from these translation functions are converted to the R factor or the correlation coefficient, enabling comparisons among the various orientations. This protocol allows the automatic examination of not only the top peaks in the rotation functions, but also those angles that produce high rotation-function values. In addition, packing of the model in the crystal is examined automatically to eliminate those solutions that have severe clashes among the molecules (see Section 13.3.5). This generalized protocol has proven more powerful than conventional methods in structure determinations (Tong, 1996*a*; Wu *et al.*, 1997). An alternative approach, examining the neighbourhood surrounding the rotation-function peaks, is also possible (Urzhumtsev & Podjarny, 1995).

Stochastic or genetic algorithms have also been used successfully to simultaneously sample the six degrees of freedom (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000).

13.3.8. The locked translation function

In the presence of noncrystallographic symmetry (NCS), locked self-rotation functions can be used to determine the orientation of the NCS elements in the crystal unit cell (Tong & Rossmann, 1990). Often, an atomic model for the monomer of the NCS assembly is available, but not a model of the entire assembly. This atomic model can be used in ordinary cross-rotation-function calculations. A more powerful technique is to use the locked cross-rotation function, which can define the orientations of all the molecules within the assembly at the same time (Tong & Rossmann, 1997). With the knowledge of the orientations, several translation searches are needed to locate the individual monomers of the assembly. For cases where the assembly has high NCS, the translation searches to locate the first few molecules may not be very successful, since the search model only represents a small portion of the diffracting power of the crystal.

A locked translation function takes into account contributions from all the monomers of the assembly at the same time (Tong, 1996*b*). It can determine the position of the monomer search model relative to the centre of the NCS assembly. With this knowledge, the entire assembly can be generated and can then be used in an ordinary translation search to locate the centre of this NCS assembly in the crystal unit cell.

Given the atomic model, X_j^0 (in Cartesian coordinates), for the monomer at a starting position and the rotation, $[F]$, that brings it into the same orientation as that of a monomer in the standard orientation, the model of the entire assembly in the standard orientation is given by

$$X_{j,m} = [I_m]([F]X_j^0 + V_0), \tag{13.3.8.1}$$

where V_0 is a translation vector and the centre of the assembly is placed at (0, 0, 0). $[I_m]$ $(m = 1, \ldots, M)$ is the set of rotation matrices for the NCS point group in the standard orientation. The correct translation vector should give rise to the maximal overlap between the self vectors within the NCS assembly and the observed Patterson map. This overlap is given by the second term of equation (13.3.3.2). The locked translation function is therefore defined as [a comparison with equation (13.3.2.9) showing that the summation over the crystallographic symmetry operators can be omitted]

$$\begin{aligned} LTF(V_0) = {} & \sum_h \sum_m (F_h^o)^2 \mid \overline{f}_{h,m}|^2 \\ & + \sum_h \sum_m \sum_{n \neq m} (F_h^o)^2 \overline{f}_{h,m}\, \overline{f}^*_{h,n} \exp\{-2\pi i h([\theta_n] - [\theta_m])V_0\}, \end{aligned} \tag{13.3.8.2}$$

where

$$\overline{f}_{h,m} = \sum_j f_j \exp(2\pi i h[\theta_m][F]X_j^0) \tag{13.3.8.3}$$

and

$$[\theta_m] = [\alpha][E][I_m]. \tag{13.3.8.4}$$

$[E]$ is the rotation matrix that brings the standard orientation to that of the assembly in the crystal and $[\alpha]$ is the de-orthogonalization matrix (Rossmann & Blow, 1962). Equation (13.3.8.2) can be evaluated indirectly by the FFT technique (Tong, 1996*b*). As with the Patterson-correlation translation function, equation (13.3.8.2) can be converted to a correlation coefficient, although the evaluation will become more time-consuming. It should be noted that equation (13.3.8.2) bears much resemblance to equation (13.3.3.2), with the interchange of the crystallographic quantities $([T_n], \overline{f}_{h,n})$ and the noncrystallographic quantities $([\theta_m], \overline{f}_{h,m})$.

13.3.9. Miscellaneous translation functions

A variety of translation functions have been developed for special purposes. For example, a function has been proposed that can determine the translational component along an (NCS) twofold axis (Rossmann & Blow, 1962).

If initial phase information is available for a crystal that possesses NCS, averaging among the NCS monomers is a powerful technique for improving the phase information (Rossmann, 1990). Accurate parameters for the NCS elements (orientation and position) are essential for this averaging process. The orientations of the NCS elements can often be determined by self-rotation functions. The positions of the NCS elements can be determined by a special translation function based on the electron-density overlap (in a spherical volume) of NCS-related monomers (Tong, 1993):

$$T([C], s_1, s_2) = \sum_h \sum_p \overline{F}_h \overline{F}_p G_{hp} \exp(-2\pi i h s_1) \exp(-2\pi i p s_2), \tag{13.3.9.1}$$

where s_1 and s_2 are the centres of two molecules related by NCS, and $[C]$ is the rotation matrix for the NCS. This equation can be used to obtain an initial estimate for the position of the NCS axis (given its orientation) (Tong, 1993), to identify positions related by the NCS (Blow *et al.*, 1964) and to refine the NCS parameters iteratively (Tong *et al.*, 1992, 1997; Tong & Rossmann, 1997).

References

Bentley, G. A. & Houdusse, A. (1992). *Some applications of the phased translation function in macromolecular structure determination. Acta Cryst.* A**48**, 312–322.

Blow, D. M., Rossmann, M. G. & Jeffery, B. A. (1964). *The arrangement of α-chymotrypsin molecules in the monoclinic crystal form. J. Mol. Biol.* **8**, 65–78.

Brünger, A. T. (1990). *Extension of molecular replacement: a new search strategy based on Patterson correlation refinement. Acta Cryst.* A**46**, 46–57.

Chang, G. & Lewis, M. (1997). *Molecular replacement using genetic algorithms. Acta Cryst.* D**53**, 279–289.

Colman, P. M., Fehlhammer, H. & Bartels, K. (1976). *Patterson search methods in protein structure determination: β-trypsin and immunoglobulin fragments*. In *Crystallographic Computing Techniques*, edited by F. R. Ahmed, K. Huml & B. Sedlacek, pp. 248–258. Copenhagen: Munksgaard.

Cowtan, K. (1998). *Modified phased translation functions and their application to molecular-fragment location. Acta Cryst.* D**54**, 750–756.

Crowther, R. A. & Blow, D. M. (1967). *A method of positioning a known molecule in an unknown crystal structure. Acta Cryst.* **23**, 544–548.

Driessen, H. P. C., Bax, B., Slingsby, C., Lindley, P. F., Mahadevan, D., Moss, D. S. & Tickle, I. J. (1991). *Structure of oligomeric βB2-crystallin: an application of the T_2 translation function to an asymmetric unit containing two dimers. Acta Cryst.* B**47**, 987–997.

Fujinaga, M. & Read, R. J. (1987). *Experiences with a new translation-function program. J. Appl. Cryst.* **20**, 517–521.

Glykos, N. M. & Kokkinidis, M. (2000). *A stochastic approach to molecular replacement. Acta Cryst.* D**56**, 169–174.

Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *A translation function combining packing and diffraction information: an application to lysozyme (high-temperature form). Acta Cryst.* A**37**, 398–406.

Hendrickson, W. A. & Ward, K. B. (1976). *A packing function for delimiting the allowable locations of crystallized macromolecules. Acta Cryst.* A**32**, 778–780.

Hirshfeld, F. L. (1968). *Symmetry in the generation of trial structures. Acta Cryst.* A**24**, 301–311.

Jogl, G., Tao, X., Xu, Y. & Tong, L. (2001). *COMO: a program for combined molecular replacement. Acta Cryst.* D**57**, 1127–1134.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Rapid automated molecular replacement by evolutionary search. Acta Cryst.* D**55**, 484–491.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software. J. Appl. Cryst.* **40**, 658–674.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Likelihood-enhanced fast translation functions. Acta Cryst.* D**61**, 458–464.

Navaza, J. (1994). *AMoRe: an automated package for molecular replacement. Acta Cryst.* A**50**, 157–163.

Navaza, J. & Vernoslova, E. (1995). *On the fast translation functions for molecular replacement. Acta Cryst.* A**51**, 445–449.

Nixon, P. E. & North, A. C. T. (1976). *Crystallographic relationship between human and hen-egg lysozymes. I. Methods for the establishment of molecular orientational and positional parameters. Acta Cryst.* A**32**, 320–325.

Rabinovich, D. & Shakked, Z. (1984). *A new approach to structure determination of large molecules by multi-dimensional search methods. Acta Cryst.* A**40**, 195–200.

Rae, A. D. (1977). *The use of structure factors to find the origin of an oriented molecular fragment. Acta Cryst.* A**33**, 423–425.

Read, R. J. (2001). *Pushing the boundaries of molecular replacement with maximum likelihood. Acta Cryst.* D**57**, 1373–1382.

Read, R. J. & Schierbeek, A. J. (1988). *A phased translation function. J. Appl. Cryst.* **21**, 490–495.

Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G. (1990). *The molecular replacement method. Acta Cryst.* A**46**, 73–82.

Rossmann, M. G. (2001). *Molecular replacement – historical background. Acta Cryst.* D**57**, 1360–1366.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *Implementation of a six-dimensional search using the AMoRe translation function for difficult molecular-replacement problems. J. Appl. Cryst.* **32**, 98–101.

Tickle, I. J. (1985). *Review of space group general translation functions that make use of known structure information and can be expanded as Fourier series.* In *Proceedings of the Daresbury Study Weekend. Molecular Replacement*, edited by P. A. Machin, pp. 22–26. Warrington: Daresbury Laboratory.

Tollin, P. (1966). *On the determination of molecular location. Acta Cryst.* **21**, 613–614.

Tollin, P. & Rossmann, M. G. (1966). *A description of various rotation function programs. Acta Cryst.* **21**, 872–876.

Tong, L. (1993). *Replace, a suite of computer programs for molecular-replacement calculations. J. Appl. Cryst.* **26**, 748–751.

Tong, L. (1996*a*). *Combined molecular replacement. Acta Cryst.* A**52**, 782–784.

Tong, L. (1996*b*). *The locked translation function and other applications of a Patterson correlation function. Acta Cryst.* A**52**, 476–479.

Tong, L., Choi, H.-K., Minor, W. & Rossmann, M. G. (1992). *The structure determination of Sindbis virus core protein using isomorphous replacement and molecular replacement averaging between two crystal forms. Acta Cryst.* A**48**, 430–442.

Tong, L., Qian, C., Davidson, W., Massariol, M.-J., Bonneau, P. R., Cordingley, M. G. & Lagacé, L. (1997). *Experiences from the structure determination of human cytomegalovirus protease. Acta Cryst.* D**53**, 682–690.

Tong, L. & Rossmann, M. G. (1990). *The locked rotation function. Acta Cryst.* A**46**, 783–792.

Tong, L. & Rossmann, M. G. (1997). *Rotation function calculations with GLRF program. Methods Enzymol.* **276**, 594–611.

Urzhumtsev, A. & Podjarny, A. (1995). *On the solution of the molecular-replacement problem at very low resolution: application to large complexes. Acta Cryst.* D**51**, 888–895.

Wu, H., Kwong, P. D. & Hendrickson, W. A. (1997). *Dimeric association and segmental variability in the structure of human CD4. Nature (London)*, **387**, 527–530.

Zhang, X.-J. & Matthews, B. W. (1994). *Enhancement of the method of molecular replacement by incorporation of known structural information. Acta Cryst.* D**50**, 675–686.

International Tables for Crystallography (2012). Vol. F, Chapter 13.4, pp. 352–363.

Chapter 13.4. Noncrystallographic symmetry averaging of electron density for molecular-replacement phase refinement and extension

M. G. Rossmann and E. Arnold

13.4.1. Introduction

Electron-density averaging for phasing crystal structures has become a widespread and nearly routine technique. High noncrystallographic symmetry (NCS) permits the solution of structures using relatively poor and even *ab initio* phasing starts, while lower NCS electron-density averaging can significantly improve initial phases obtained by techniques such as isomorphous replacement, anomalous scattering, or molecular replacement. Implicit in any averaging is solvent flattening in the regions not available for NCS averaging. Indeed, if all parts of the unit cell were consistent with the NCS, the symmetry would be crystallographic and valid throughout the crystal lattice.

A number of generalized averaging programs and software packages have been developed for macromolecular crystal structure analyses. Ease of use, coupled with relatively convenient definition of molecular envelopes, as well as enormous advances in computer technology, have facilitated the application of symmetry averaging to a diverse set of crystallographic problems. Averaging of separate domains in multidomain protein structures that can be divided into segments and averaging among multiple crystal forms is becoming increasingly common.

Extension of phases to higher resolution by symmetry averaging of electron density, coupled with solvent flattening, has been applied to numerous problems. The power of phase extension has been especially impressive in many cases involving high NCS, such as icosahedral virus structures. The overall power for phase improvement of averaging, combined with other density-modification techniques, such as solvent levelling, has been found to depend upon the degree of NCS, the solvent content of the crystals, and the quality and completeness of experimental data. Similar averaging methodology can be used for structure analysis by other imaging techniques, such as electron microscopy.

This chapter discusses the underlying principles of electron-density averaging for macromolecular crystallographic phase improvement and describes procedures for computer implementation of these techniques.

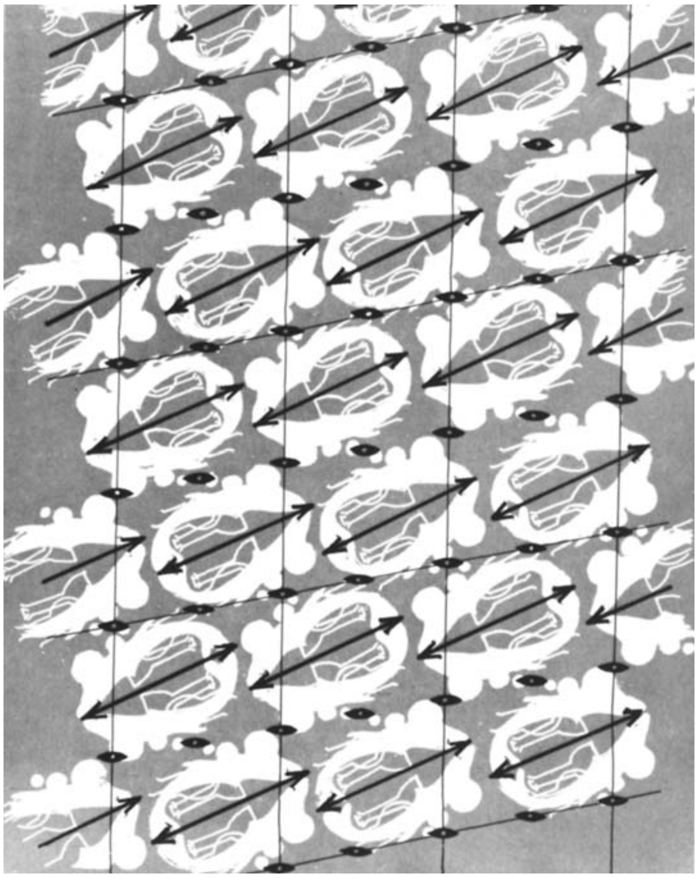


Figure 13.4.2.1
The two-dimensional periodic design shows crystallographic twofold axes perpendicular to the page and local noncrystallographic rotation axes in the plane of the paper (design by Audrey Rossmann). [Reprinted with permission from Rossmann (1972). Copyright (1972) Gordon & Breach.]

13.4.2. Noncrystallographic symmetry (NCS)

Crystallographic symmetry is valid for the infinite crystal lattice. Any crystallographic symmetry element relates all points within the crystal to equivalent points elsewhere. In contrast, an NCS operator is valid only locally within a finite volume (Fig. 13.4.2.1); if a periodic structure is superimposed on itself after operation with an NCS operator, it will superimpose only within the envelope[1] defining the limits of the local symmetry.

A product of superimposed periodic structures will be nonperiodic, containing only the point symmetry of the noncrystallographic operators (Fig. 13.4.2.2). This fact can frequently be used to select a molecular envelope that was not obvious prior to noncrystallographic averaging [see *e.g.* Buehner *et al.* (1974) or Lin *et al.* (1986)]. Although no knowledge of the crystallographic envelope is needed for this first averaging, it is necessary to determine it for the averaged molecular structure within the crystallographic cell to permit Fourier back-transformation.

There must be space between the confining envelopes governed by the local symmetry. Only the crystallographic symmetry is valid within this space. In the limit, when this space has diminished to zero, the local symmetry will have become a true crystallographic operator.

The definition of NCS can be extended to symmetry that relates similar objects in different crystal lattices. An operation that relates an object in one lattice to an equivalent object in another lattice will apply only to the chosen objects in each lattice. Beyond the confines of the chosen objects, there will be no coincidence of pattern.

Two kinds of NCS elements may be defined: proper and improper. The former satisfies a closed point group [*e.g.* a 17-fold rotation as occurs in tobacco mosaic virus disk protein

---

[1] In this chapter, 'envelope' will be used to describe the external surface or boundary of a molecule, while 'mask' will be used to denote the three-dimensional distribution of grid points that have been assigned within the molecular surface.

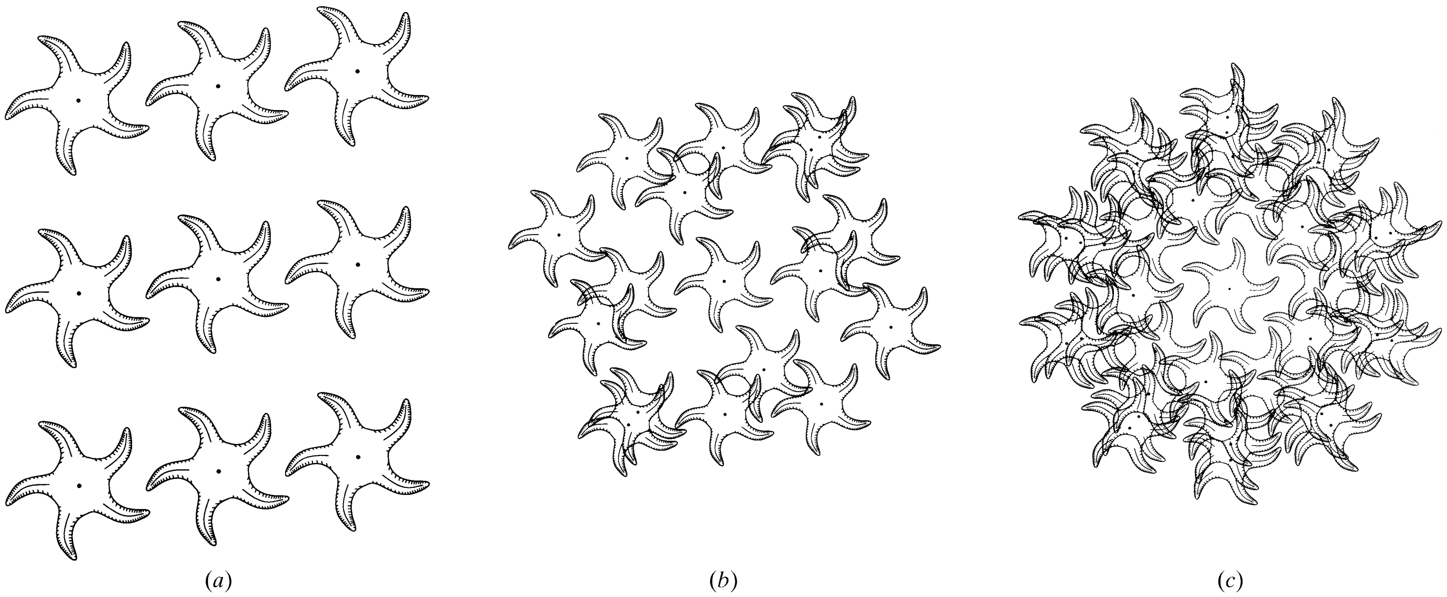


Figure 13.4.2.2
(*a*) NCS in a triclinic cell. (*b*) Superposition of the pattern in (*a*) on itself after operation with the noncrystallographic fivefold axis. (*c*) Superposition of the pattern in (*a*) on itself after a rotation of one-fifth, two-fifths, three-fifths and four-fifths. Note that the sum or product of periodic patterns is aperiodic and in (*c*) has the point symmetry of the noncrystallographic operation. [Reprinted with permission from Rossmann (1990).]

(Champness *et al.*, 1976)]. Here, it does not matter whether a rotation axis is applied right- or left-handedly; the result is indistinguishable. On the other hand, the relationship between different molecules in a crystallographic asymmetric unit is unlikely to be a closed point group. Thus, a rotation in one direction (followed by a translation) might achieve superposition of the two molecules, while a rotation in the opposite direction would not. This is called an improper NCS operator. An operation which takes a molecule in one unit cell to that in another unit cell (initially, the cells are lined up with, say, their orthogonalized *a*, *b* and *c* axes parallel) must equally be an improper rotation.

The position in space of a noncrystallographic rotation symmetry operator can be arbitrarily assigned. The rotation operation will orient the two molecules similarly. A subsequent translation, whose magnitude depends upon the location of the NCS operator, will always be able to superimpose the molecules (Fig. 13.4.2.3). Nevertheless, it is possible to select the position of the NCS axis such that the translation is a minimum, and that will occur when the translation is entirely parallel to the noncrystallographic rotation axis.

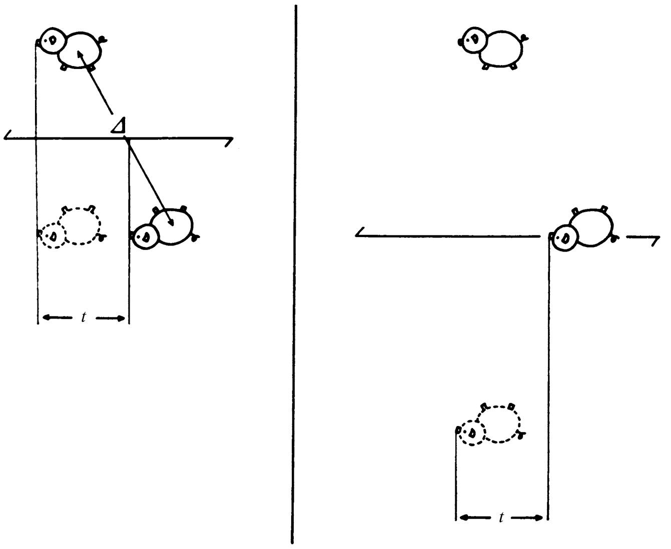


Figure 13.4.2.3
The position of the twofold rotation axis which relates the two piglets is completely arbitrary. The diagram on the left shows the situation when the translation is parallel to the rotation axis. The diagram on the right has an additional component of translation perpendicular to the rotation axis, but the component parallel to the axis remains unchanged. [Reprinted with permission from Rossmann *et al.* (1964).]

The position of an NCS axis, like everything else in the unit cell, must be defined with respect to a selected origin. Consider the noncrystallographic rotation defined by the 3×3 matrix [C]. Then, if the point $\mathbf{x}$ is rotated to $\mathbf{x}'$ (both defined with respect to the selected origin and axial system),

$$\mathbf{x}' = [\mathrm{C}]\mathbf{x} + \mathbf{d},$$

where $\mathbf{d}$ is a three-dimensional vector which expresses the translational component of the NCS operation. The magnitude of the components of $\mathbf{d}$ is quite arbitrary unless the position of the rotation axis is defined. If the rotation axis represents a proper NCS element, there will exist a point $\mathbf{x}$ on the rotation axis, when positioned to eliminate translation, such that it is rotated onto $\mathbf{x}'$. It follows that for such a point

$$\mathbf{x} = [\mathrm{C}]\mathbf{x} + \mathbf{d},$$

from which $\mathbf{d}$ can be determined if the position of the molecular centre is known. Note that $\mathbf{d} = 0$ if, and only if, the noncrystallographic rotation axis passes through the crystallographic origin.

The presence of proper NCS in a crystal can help phase determination considerably. Consider, for example, a tetramer with 222 symmetry. It is not necessary to define the chemical limits of any one polypeptide chain as the NCS is true everywhere within the molecular envelope and the boundaries of the polypeptide chain are irrelevant to the geometrical considerations. The electron density at every point within the molecular envelope (which itself must have 222 symmetry) can be averaged among all four 222-related points without any chemical knowledge of the configuration of the monomer polypeptide. On the other hand, if there is only improper NCS, then the envelope must define the limits of one noncrystallographic asymmetric unit, although the crystallographic asymmetric unit contains two or more such units.

13.4.3. Phase determination using NCS

The molecular replacement method [*cf.* Rossmann & Blow (1962); Rossmann (1972, 1990); Argos & Rossmann (1980); Rossmann & Arnold (2001)] is dependent upon the presence of NCS, whether it relates objects within one crystal lattice or between crystal lattices. The NCS rotational relationship in real space is exactly mimicked in reciprocal space. Local symmetry in real space has the equivalent effect of rotating a reciprocal lattice onto itself or another (with origins coincident), such that the integral reciprocal-lattice points of one reciprocal space coincide with non-integral reciprocal-lattice positions in the other. As the reciprocal lattice samples the Fourier transform of a molecule only at finite and integral reciprocal-lattice points, the effect of an NCS operation is to permit sampling of the molecular transform at intermediate non-integral reciprocal-lattice positions. If such sampling occurs frequently enough, it will constitute a plot of the continuous transform of the molecule and, hence, amount to a structure determination.

Whenever a molecule exists more than once either in the same unit cell or in different unit cells, then error in the molecular electron-density distribution due to error in phasing can be reduced by averaging the various molecular copies. The number of such copies, N, is referred to as the noncrystallographic redundancy. As the NCS is, by definition, only local (often pertaining to a particular molecular centre), there are holes and gaps between the averaged density, which presumably are solvent space between molecules. Thus, the electron density can be improved both by averaging electron density and by setting the density between molecules to a low, constant value ('solvent flattening'). Phases calculated by Fourier back-transforming the improved density should be more accurate than the original phases. Hence, the observed structure amplitudes (suitably weighted) can be associated with the improved phases, and a new and improved map can be calculated. This, in turn, can again be averaged until convergence has been reached and the phases no longer change. In addition, the back-transformed map can be used to compute phases just beyond the extremity of the resolution of the terms used in the original map. The resultant amplitudes will not be zero because the map had been modified by averaging and solvent flattening. Thus, phases can be gradually extended and improved, starting from a very low resolution approximation to the molecular structure. This procedure was first implemented in reciprocal space (Rossmann & Blow, 1963; Main, 1967; Crowther, 1969) and then, more recently, in real space (Bricogne, 1974, 1976; Johnson, 1978; Jones, 1992; Rossmann *et al.*, 1992). More recently still, there has been an attempt to reproduce the very successful real-space procedure in reciprocal space (Tong & Rossmann, 1995).

Early examples of such a procedure for phase improvement are the structure determinations of deoxyhaemoglobin (Muirhead *et al.*, 1967), α-chymotrypsin (Matthews *et al.*, 1967), lobster glyceraldehyde-3-phosphate dehydrogenase (Buehner *et al.*, 1974), hexokinase (Fletterick & Steitz, 1976), tobacco mosaic virus disk protein (Champness *et al.*, 1976; Bloomer *et al.*, 1978), the influenza virus haemagglutinin spike (Wilson *et al.*, 1981), tomato bushy stunt virus (Harrison *et al.*, 1978) and southern bean mosaic virus (Abad-Zapatero *et al.*, 1980). Early examples of phase extension, using real-space electron-density averaging, were the study of glyceraldehyde-3-phosphate dehydrogenase (Argos *et al.*, 1975), satellite tobacco necrosis virus (Nordman, 1980), haemocyanin (Gaykema *et al.*, 1984), human rhinovirus 14 (Rossmann *et al.*, 1985) and poliovirus (Hogle *et al.*, 1985). Since then, this method has been used in numerous virus structure determinations, with the phase extension being initiated from ever lower resolution.

A once-popular computer program for real-space averaging was written by Gerard Bricogne (1976). Another program has been described by Johnson (1978). Both programs were based on a double-sorting procedure. Bricogne (1976) had suggested that, with interpolation between grid points using linear polynomials, it was necessary to sample electron density at grid intervals finer than one-sixth of the resolution limit of the Fourier terms that were used in calculating the map. With the availability of more computer memory, it was possible to store much of the electron density, thus avoiding time-consuming sorting operations (Hogle *et al.*, 1985; Luo *et al.*, 1989). Simultaneously, the storage requirements could be drastically reduced by using interpolation with quadratic polynomials. While the latter required a little extra computation time, this was far less than what would have been needed for sorting. Furthermore, it was found that Bricogne's estimate for the fineness of the map storage grid was too pessimistic, even for linear interpolation, which works well to about 1/2.5 of the resolution limit of the map.

In addition to changes in strategy brought about by computers with much larger memories, experience has been gained in program requirements for real-space averaging for phase determination (Dodson *et al.*, 1992). Here we give a general procedure for electron-density averaging.

13.4.4. The *p*- and *h*-cells

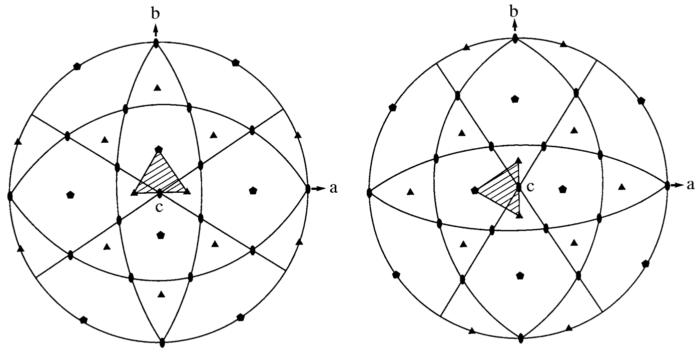
It is useful to define two types of unit cells.

(1) The 'p-cell' is the unit cell of the unknown crystal structure and is associated with fractional coordinates $\mathbf{y}$ and unit-cell vectors $\mathbf{a}_p$, $\mathbf{b}_p$, $\mathbf{c}_p$.

(2) The 'h-cell' is the unit cell with respect to which the noncrystallographic axes of the molecule (or particle) are to be defined in a standard orientation and is associated with fractional coordinates $\mathbf{x}$ and unit-cell vectors $\mathbf{a}_h$, $\mathbf{b}_h$, $\mathbf{c}_h$.

Since the averaged molecule is to be placed into all crystallographically related positions in the p-cell, it is essential to know the envelope that encloses a single molecule. Care must be taken that the envelopes from neighbouring molecules in the p-cell do not overlap. The remaining space between the limits of the envelopes of the variously placed molecules in the p-cell can be taken to be solvent and, hence, flattened, a useful physical assumption for helping phase determination.

The h-cell must be chosen to be at least as large as the largest dimension of the molecule. In general, it is convenient to define the h-cell with $a_h = b_h = c_h$ and $\alpha = \beta = \gamma = 90^\circ$, while placing the molecular centre at $(\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. For example, if the molecule is a viral particle with icosahedral symmetry, the standard orientation can be defined by placing the twofold axes to correspond to the h-cell unit-cell axes, a procedure which can be done in one of two ways (Fig. 13.4.4.1). It will be necessary to know how the molecule (or particle) in the h-cell is related to the 'reference' molecule in the p-cell. The known p-cell crystallographic symmetry then permits the complete construction of the p-cell structure from whatever is the current h-cell electron-density representation of the molecule.



(*a*) (*b*)

Figure 13.4.4.1
Stereographic projections showing alternative definitions of the 'standard orientation' of an icosahedron in the *h*-cell. Icosahedral axes are placed parallel to the cell axes. Limits of a noncrystallographic asymmetric unit are shaded, representing 1/60th of the volume of an object with icosahedral symmetry. [Reproduced with permission from Rossmann *et al.* (1992).]

The *h*-cell is used to represent the density of a molecule in the standard orientation obtained by averaging all the noncrystallographic units in the *p*-cell. While density within a specific molecule will tend to be reinforced by the averaging procedure, the density outside the molecular boundaries will tend to be diminished. Thus, by averaging into the *h*-cell, the molecular envelope is revealed automatically. Indeed, the greater the NCS, the greater the clarity of the molecular boundary. Hence, the averaged molecule in the *h*-cell can be used to define a molecular mask in the *p*-cell automatically.

Averaging into the *h*-cell is also useful for displaying the molecule in a standard orientation (*i.e.* obtaining the electron-density distribution on skew planes). Thus, it is possible to display the molecule, for instance, with sections perpendicular to a molecular twofold axis, and to position the molecular symmetry axes accurately. From this, it is then easy to define the limits of the molecular asymmetric unit (Fig. 13.4.4.1). Hence, it is possible to save a great deal of computing time by evaluating the electron density in the *h*-cell only at those grid points within and immediately surrounding the noncrystallographic asymmetric unit.

13.4.5. Combining crystallographic and noncrystallographic symmetry

Transformations will now be described which relate noncrystallographically related positions distributed among several fragmented copies of the molecule in the asymmetric unit of the *p*-cell and between the *p*-cell and the *h*-cell.

13.4.5.1. General considerations

Let $\mathbf{Y}$ and $\mathbf{X}$ be position vectors in a Cartesian coordinate system whose components have dimensions of length, in the *p*- and *h*-cells, which utilize the same origin as the fractional coordinates, $\mathbf{y}$ and $\mathbf{x}$, respectively. Let $[\beta_p]$ and $[\alpha_h]$ be 'orthogonalization' and 'de-orthogonalization' matrices in the *p*- and *h*-cells, respectively (Rossmann & Blow, 1962). Then

$$\begin{array}{lll} \mathbf{Y} = [\beta_p]\mathbf{y} & \text{and} & \mathbf{x} = [\alpha_h]\mathbf{X}, \\ [\alpha_p] = [\beta_p]^{-1} & \text{and} & [\alpha_h] = [\beta_h]^{-1}. \end{array} \qquad (13.4.5.1)$$

Thus, for instance, $[\alpha_h]$ denotes a matrix that transforms a Cartesian set of unit vectors to fractional distances along the unit-cell vectors $\mathbf{a}_h$, $\mathbf{b}_h$, $\mathbf{c}_h$.

Let the Cartesian coordinates $\mathbf{Y}$ and $\mathbf{X}$ be related by the rotation matrix $[\omega]$ and the translation vector $\mathbf{D}$ such that

$$\mathbf{X} = [\omega]\mathbf{Y} + \mathbf{D}. \qquad (13.4.5.2)$$

If the molecules are to be averaged among different unit cells, then each *p*-cell must be related to the standard *h*-cell orientation by a different $[\omega]$ and $\mathbf{D}$. Then, from (13.4.5.1) and (13.4.5.2)

$$\mathbf{X} = [\omega][\beta_p]\mathbf{y} + \mathbf{D}. \qquad (13.4.5.3)$$

Now, if $[\omega]$ represents the rotational relationship between the 'reference' molecule, $m = 1$, in the *p*-cell with respect to the *h*-cell, then from (13.4.5.3)

$$\mathbf{X} = [\omega][\beta_p]\mathbf{y}_{m=1} + \mathbf{D},$$

where $\mathbf{y}_m$ refers to the fractional coordinates of the *m*th molecule in the *p*-cell.

Assuming there is only one molecule per asymmetric unit in the *p*-cell, let the *m*th molecule in the *p*-cell be related to the reference molecule by the crystallographic rotation $[\mathrm{T}_m]$ and translational operators $\mathbf{t}_m$, such that

$$\mathbf{y}_m = [\mathrm{T}_m]\mathbf{y}_{m=1} + \mathbf{t}_m. \qquad (13.4.5.4)$$

For convenience, all translational components will initially be neglected in the further derivations below, but they will be reintroduced in the final stages. Hence, from (13.4.5.3) and (13.4.5.4)

$$\mathbf{X} = \{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_m. \qquad (13.4.5.5)$$

Further, if $\mathbf{X}_n$ refers to the *n*th subunit within the molecule in the *h*-cell, and similarly if $\mathbf{y}_{m,n}$ refers to the *n*th subunit within the *m*th molecule of the *p*-cell, then from (13.4.5.5)

$$\mathbf{X}_n = \{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_{m,n}. \qquad (13.4.5.6)$$

Finally, the rotation matrix $[\mathrm{R}_n]$ is used to define the relationship among the N ($N = 2$ for a dimer, 4 for a 222 tetramer, 60 for an icosahedral virus *etc.*) noncrystallographic asymmetric units of the molecule within the *h*-cell. Then

$$\mathbf{X}_n = [\mathrm{R}_n]\mathbf{X}_{n=1}. \qquad (13.4.5.7)$$

13.4.5.2. Averaging with the p-cell

Consider averaging the density at N noncrystallographically related points in the *p*-cell and replacing that density into the *p*-cell. By substituting for $\mathbf{X}_n$ and $\mathbf{X}_{n=1}$ in (13.4.5.7) and using (13.4.5.6),

$$\{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_{m,n} = [\mathrm{R}_n]\{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_{m,n=1},$$

or

$$\mathbf{y}_{m,n} = \{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}^{-1} \times [\mathrm{R}_n]\{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_{m,n=1}. \qquad (13.4.5.8)$$

Now set

$$\begin{aligned} [\mathrm{E}_{m,n}] &= \{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}^{-1}[\mathrm{R}_n]\{[\omega][\beta_p][\mathrm{T}_m^{-1}]\} \\ &= [\mathrm{T}_m][\alpha_p][\omega^{-1}][\mathrm{R}_n][\omega][\beta_p][\mathrm{T}_m^{-1}], \end{aligned} \qquad (13.4.5.9)$$

giving

$$\mathbf{y}_{m,n} = [\mathrm{E}_{m,n}]\mathbf{y}_{m,n=1} + \mathbf{e}_{m,n}, \qquad (13.4.5.10)$$

where $\mathbf{e}_{m,n}$ is the corresponding translational element. Note that multiplication by $[\mathrm{E}_{m,n}]$ thus corresponds to the following sequence of transformations: (1) placing all the crystallographically related subunits into the reference orientation with $[\mathrm{T}_m^{-1}]$; (2) ‘orthogonalizing’ the coordinates with $[\beta_p]$; (3) rotating the coordinates into the *h*-cell with $[\omega]$; (4) rotating from the reference subunit of the molecule of the *h*-cell with $[\mathrm{R}_n]$; (5) rotating these back into the *p*-cell with $[\omega^{-1}]$; (6) ‘de-orthogonalizing’ in the *p*-cell with $[\alpha_p]$; and (7) placing these back into each of the *M* crystallographic asymmetric units of the *p*-cell with $[\mathrm{T}_m]$.

The translational elements, $\mathbf{e}_{m,n}$, can now be evaluated. Let $\mathbf{s}_{p,m}$ be the fractional coordinates of the centre (or some arbitrary position) of the *m*th molecule in the *p*-cell; hence, $\mathbf{s}_{p,m=1}$ denotes the molecular centre position of the reference molecule in the *p*-cell. If $\mathbf{s}_{p,m}$ is at the intersection of the molecular rotation axes, then it will be the same for all *n* molecular asymmetric units. Therefore, it follows from (13.4.5.10) that

$$\mathbf{e}_{m,n} = \mathbf{s}_{p,m} - [\mathrm{E}_{m,n}]\mathbf{s}_{p,m=1}, \qquad (13.4.5.11a)$$

or

$$\mathbf{y}_{m,n} = [\mathrm{E}_{m,n}]\mathbf{y}_{m,n=1} + (\mathbf{s}_{p,m} - [\mathrm{E}_{m,n}]\mathbf{s}_{p,m=1}). \qquad (13.4.5.11b)$$

Equation (13.4.5.11*b*) can be used to find all the *N* noncrystallographic asymmetric units within the crystallographic asymmetric unit of the *p*-cell. Thus, this is the essential equation for averaging the density in the *p*-cell and replacing it into the *p*-cell.

13.4.5.3. *Averaging the p-cell and placing the results into the h-cell*

Consider averaging the density at *N* noncrystallographically related points in the *p*-cell and placing that result into the *h*-cell. From (13.4.5.7), multiplying by $[\alpha_h]$,

$$[\alpha_h][\mathbf{X}_{n=1}] = [\alpha_h][\mathrm{R}_n^{-1}]\mathbf{X}_n.$$

From (13.4.5.1) and (13.4.5.6),

$$\mathbf{x}_{n=1} = [\alpha_h][\mathrm{R}_n^{-1}]\{[\omega][\beta_p][\mathrm{T}_m^{-1}]\}\mathbf{y}_{m,n}. \qquad (13.4.5.12)$$

Since it is only necessary to place the reference molecule of the *p*-cell into the *h*-cell, it is sufficient to consider the case when $m = 1$, in which case $[\mathrm{T}_m^{-1}]$ is the identity matrix [I]. It then follows, by inversion, that

$$\begin{aligned}\mathbf{y}_{m=1,n} &= \{[\omega][\beta_p]\}^{-1}[\mathrm{R}_n][\beta_h]\mathbf{x}_{n=1}\\ &= [\alpha_p][\omega^{-1}][\mathrm{R}_n][\beta_h]\mathbf{x}_{n=1},\end{aligned}$$

which corresponds to: (1) ‘orthogonalizing’ the *h*-cell fractional coordinates with $[\beta_h]$; (2) rotating into the *n*th noncrystallographic unit within the molecule using $[\mathrm{R}_n]$; (3) rotating into the *p*-cell with $[\omega^{-1}]$; and (4) ‘de-orthogonalizing’ into fractional *p*-cell coordinates with $[\alpha_p]$.

Now, if $\mathbf{s}_h$ is the molecular centre in the *h*-cell (usually $\frac{1}{2}, \frac{1}{2}, \frac{1}{2}$), then

$$\mathbf{y}_{m=1,n} = [\mathrm{E}'_{m=1,n}]\mathbf{x} + (\mathbf{s}_{p,m=1} - [\mathrm{E}'_{m=1,n}]\mathbf{s}_h),$$

and

$$[\mathrm{E}'_{m=1,n}] = [\alpha_p][\omega^{-1}][\mathrm{R}_n][\beta_h]. \qquad (13.4.5.13)$$

Equation (13.4.5.13) determines the position of the *N* noncrystallographically related points $\mathbf{y}_{m=1,n}$ in the *p*-cell whose average value is to be placed at $\mathbf{x}$ in the *h*-cell.

13.4.6. Determining the molecular envelope

Various techniques are available for determining the molecular envelope within which density can be averaged and outside of which the solvent can be flattened.

(1) By assumption of a simple geometric shape, such as a sphere. This is frequently used for icosahedral viruses.

(2) By manual inspection of a poor electron-density map which, nevertheless, gives some guidance as to the molecular boundaries. A variety of interactive graphical programs are available to help define the molecular boundary.

(3) By use of a homologous structure or other information, such as a cryo-electron-microscopy (cryo-EM) reconstruction at low resolution. The information about a homologous structure may be either in the form of an electron-density grid or, often more conveniently, as an atomic model.

(4) By inspection of an averaged map which should have weaker density beyond the limits of the molecular boundary where the NCS is no longer true.

Procedures (2) and (3) are advisable when the NCS redundancy is low. Procedure (4) works well when the NCS redundancy is four or higher. The crystallographic asymmetric unit is likely to contain bits and pieces of molecules centred at various positions in the unit cell and neighbouring unit cells. Therefore, it is necessary to associate each grid point within the *p*-cell crystallographic asymmetric unit to a specific molecular centre or to solvent.

If the molecular-boundary assignments are to be made automatically, then the following procedure can be used. The number, *M*, of such molecules can be estimated by generating all centres, derived from the given position of the centre for the reference molecule, $\mathbf{s}_{p,n=1}$, and then determining whether a molecule of radius R_{out} would impinge on the crystallographic asymmetric unit within the defined boundaries. Here, R_{out} is a liberal estimate of the molecular radius. The corresponding rotation matrices $[\mathrm{E}_{m,n}]$ and translation vectors $\mathbf{e}_{m,n}$ can then be computed from (13.4.5.9) and (13.4.5.11*a*).

Any grid point whose distance from all *M* centres is greater than R_{out} can immediately be designated as being in the solvent region. For other grid points, it is necessary to examine the corresponding *h*-cell density. From (13.4.5.12), it follows that (setting $n = 1$)

$$\mathbf{x} = [\mathrm{E}''_{m,n=1}]\mathbf{y}_m + (\mathbf{s}_h - [\mathrm{E}''_{m,n=1}]\mathbf{s}_{p,m}),$$

where

$$[\mathrm{E}''_{m,n=1}] = [\alpha_h][\mathrm{R}_n^{-1}][\omega][\beta_p][\mathrm{T}_m^{-1}] \qquad (13.4.6.1)$$

(*n* can be set to 1, since the *h*-cell presumably contains an averaged molecular electron density, in which case it does not matter which molecular asymmetric unit is referenced). Thus, (13.4.6.1) can be used to determine the electron density at $\mathbf{y}_m$ by inspecting the corresponding interpolated density, $\rho(\mathbf{x})$, at $\mathbf{x}$ in the *h*-cell. Transfer of the electron density, $\rho(\mathbf{x})$, from the *h*-cell to the *p*-cell using (13.4.6.1) is often useful to obtain an initial structure. However, to determine a suitable mask, it is useful to evaluate a modified electron density, $\langle\rho(\mathbf{x})\rangle$, (see below) for the grid points immediately around $\mathbf{x}$ in the *h*-cell.

A variable parameter ‘CRIT’ can be specified to establish the distribution of grid points that are within the molecular envelope. When the modified electron density, $\langle\rho(\mathbf{x})\rangle$, is less than CRIT, the corresponding grid point at $\mathbf{y}$ is assumed to be in solvent. Otherwise, when $\langle\rho(\mathbf{x})\rangle$ exceeds CRIT, the grid point at $\mathbf{y}$ is assigned to that molecule which has the largest $\langle\rho(\mathbf{x})\rangle$. If the percentage of grid points which might be assigned to more than one molecule is large (say, greater than 1% of the total number of grid points), it probably signifies that the value of CRIT is too low, that the molecular boundary is far from clear, or that the

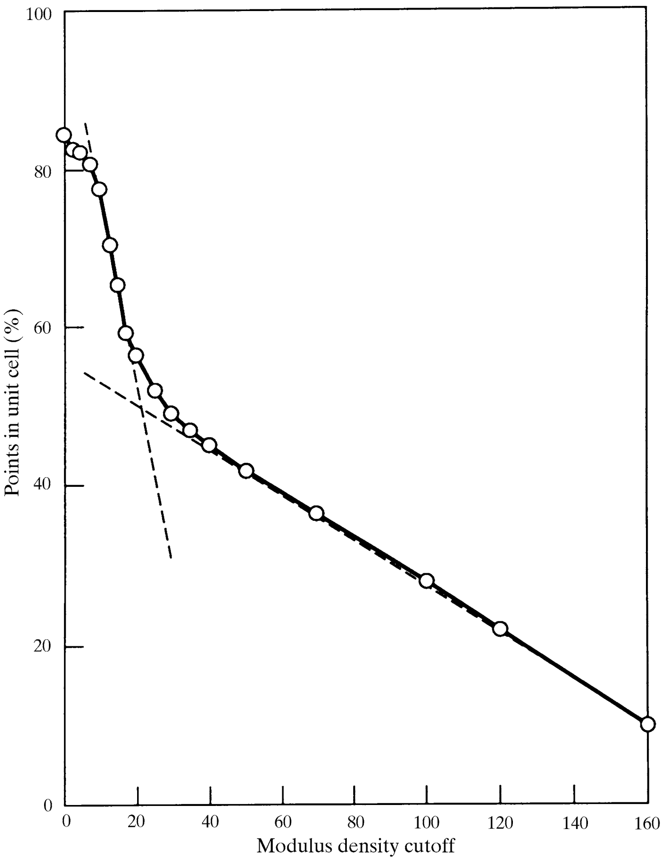


Figure 13.4.6.1
The volume of the molecular mask expressed as a percentage of the volume of the p-cell asymmetric unit, as determined by the density cutoff in the h-cell. When the modulus of the density cutoff is decreased to less than the mean smeared electron density within the protein, the mask volume increases rapidly. Intersection of the tangents suggests the most appropriate density cutoff value for mask generation. [Reproduced with permission from McKenna, Xia, Willingmann, Ilag & Rossmann (1992).]

function used to define $\langle\rho(\mathbf{x})\rangle$ was badly chosen (Fig. 13.4.6.1). Grid points outside the molecular envelope can be set to the average solvent density.

An essential criterion for the molecular envelope is that it obeys the noncrystallographic point-group symmetry. If the original h-cell electron density already possesses the molecular symmetry (*e.g.* icosahedral 532, 222 *etc.*), then the p-cell mask should also have that symmetry. However, if the mask boundaries were chosen manually, masks from different molecular centres might be in conflict and have local errors in the correct molecular symmetry. Such errors can be corrected by reimposing the noncrystallographic point-group symmetry on the p-cell mask. This can be conveniently achieved by setting the density at each grid point that was considered within the molecular envelope to a value of 100, and all other grid points to a density of zero. If the resultant density is averaged using the same routine as is used for averaging the actual electron density of the molecule, then the average density will remain 100 if the interpolated density is 100 at all noncrystallographically related points. However, if the original grid point is near the edge of the mask, finding the density at symmetry-related points may involve interpolation between density at level 100 and at level 0, giving an averaged density of less than 100. Hence, any grid point whose averaged density is below some criterion should be attributed to solvent.

Other improvements to mask generation were discussed by Rossmann *et al.* (1992). In any event, the molecular-envelope definition should be periodically re-examined after a suitable number of electron-density-averaging cycles.

Table 13.4.7.1
Mean root-mean-square scatter between noncrystallographically related points

Example taken from φX174 structure determination. $\langle\rho_8\rangle$ is proportional to the mean density (e Å^{-3}) based on eight-point interpolation; n is number of grid points with $\langle\rho_8\rangle$ in a given range; $\langle\sigma(\rho_8)\rangle$ is the root-mean-square deviation from ρ_8 among noncrystallographic asymmetric points averaged over all points in the mask.

	Density derived from an electron microscopy image at 25 Å resolution		Density derived from a 3.3 Å crystal structure	
$\langle\rho_8\rangle$	n	$\langle\sigma(\rho_8)\rangle$	n	$\langle\sigma(\rho_8)\rangle$
−375 to −325	1	44.7	0	0.0
−325 to −275	16	44.4	0	0.0
−275 to −225	22	39.5	41	31.4
−225 to −175	81	34.9	3493	25.5
−175 to −125	299	34.7	65049	20.5
−125 to −75	1119	33.1	290025	17.7
−75 to −25	16617	34.7	661386	15.0
−25 to 25	33818	46.9	1,016274	12.8
25 to 75	6008	31.9	344620	16.3
75 to 125	4512	32.0	215036	18.9
125 to 175	3050	32.1	146690	22.1
175 to 225	1562	32.6	58155	26.3
225 to 275	542	33.4	6032	32.2
275 to 325	213	35.6	227	40.6
325 to 375	33	34.7	9	46.8

13.4.7. Finding the averaged density

Electron density can be averaged (1) among the N NCS-related molecules in the p-cell (the real crystal unit cell), thus creating a new and improved map of the p-cell; (2) among the N NCS-related molecules in the p-cell and placing the results into a standard orientation in the h-cell; or (3) among the N NCS-related molecules in different unit cells and placing the results back into the original different unit cells or into a standard h-cell. Before averaging commences, the $M \times N$ matrices $[\mathrm{E}_{m,n}]$ and translation vectors $\mathbf{e}_{m,n}$ must be evaluated [see (13.4.5.9) and (13.4.5.11*a*)]. Here, N is the noncrystallographic redundancy and M is the number of molecules that impinge on the crystallographic asymmetric unit of the p-cell. Associated with each grid point in the p-cell asymmetric unit will be (1) the value of m designating which molecular centre is to be associated with that grid point (a special value of m is for solvent) and (2) the p-cell electron density at that point.

The grid points within the asymmetric unit are then examined one at a time. If the grid point is within the mask, it is averaged among the N noncrystallographically related equivalent positions belonging to molecule m. If the grid point is solvent, the density can be set to the average solvent density.

The N noncrystallographically equivalent non-integral grid points can be computed from (13.4.5.11*a*). Some of these will lie outside the crystallographic asymmetric unit. These will, therefore, have to be operated on by unit-cell translations and crystallographic symmetry operations to bring them back into the asymmetric unit before the corresponding interpolated density can be calculated.

Averaging into the h-cell can be done by a procedure similar to averaging in the p-cell, except that the rotation and translation matrices are given by (13.4.5.13). Furthermore, no mask is required as all the averaging into the h-cell (from p-cell electron

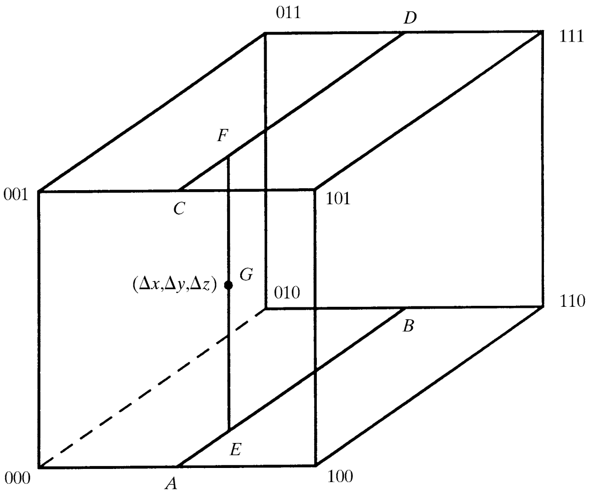


Figure 13.4.8.1
Interpolation box for finding the approximate electron density at $G(\Delta x, \Delta y, \Delta z)$, given the eight densities at the corners of the box. The interpolated value can be built up by first using interpolations to determine the densities at A, B, C and D. A second linear interpolation then determines the density at E (from densities at A and B) and at F (from densities at C and D). The third linear interpolation determines the density at G from the densities at E and F. [Reproduced with permission from Rossmann *et al.* (1992).]

density) can be done with respect to the reference molecule centred at $\mathbf{s}_{p,\,m=1}$ in the p-cell. Each grid point is taken in turn in the h-cell. The electron density at any grid point that is further away from $\mathbf{s}_{\rm h}$ than from $R_{\rm out}$ is set to zero. Other grid-point positions are expanded into the N equivalent positions in the p-cell surrounding $\mathbf{s}_{p,\,m=1}$. The interpolated density is then found, averaged over the N equivalent positions, and stored at the original h-cell grid point in successive sections, in the same way as in the p-cell averaging. As in averaging within the p-cell, a record is kept of $\langle\sigma(\rho)\rangle$ as a function of $\langle\rho(\mathbf{x})\rangle$ (Table 13.4.7.1). In general, the local NCS is valid only within the molecule. Hence, the h-cell density will show the molecular envelope and can be used to recompute an improved p-cell density mask. The rate of build up of signal within the molecule should be roughly proportional to N, while the rate outside the molecule should be proportional to about $N^{1/2}$.

13.4.8. Interpolation

Some thought must go into defining the size of the grid interval. Shannon's sampling theorem shows that the grid interval must never be greater than half the limiting resolution of the data. Thus, for instance, if the limiting resolution is 3 Å, the grid intervals must be smaller than 1.5 Å. Clearly, the finer the grid interval, the more accurate the interpolated density, but the computing time will increase with the inverse cube of the size of the grid step. Similarly, if the grid interval is fine, less care and fewer points can be used for interpolation, thus balancing the effect of the finer grid in terms of computing time. In practice, it has been found that an eight-point interpolation (as described below) can be used, provided the grid interval is less than 1/2.5 of the resolution (Rossmann *et al.*, 1992). Other interpolation schemes have also been used (*e.g.* Bricogne, 1976; Nordman, 1980; Hogle *et al.*, 1985; Bolin *et al.*, 1993).

A straightforward 'linear' interpolation can be discussed with reference to Fig. 13.4.8.1 (in mathematical literature, this is called a trilinear approximation or a tensor product of three one-dimensional linear interpolants). Let G be the position at which the density is to be interpolated, and let this point have the fractional grid coordinates Δx, Δy, Δz within the box of surrounding grid points. Let 000 be the point at $\Delta x = 0$, $\Delta y = 0$, $\Delta z = 0$. Other grid points will then be at 100, 010, 001 *etc.*, with the point diagonally opposite the origin at 111.

The density at A (between 000 and 100) can then be approximated as the value of the linear interpolant of ρ_{000} and ρ_{100}:

$$\rho(A) \cong \rho_A = \rho_{000} + (\rho_{100} - \rho_{000})\Delta x.$$

Similar expressions for $\rho(B)$, $\rho(C)$ and $\rho(D)$ can also be written. Then, it is possible to calculate an approximate density at E from

$$\rho(E) \cong \rho_E = \rho_A + (\rho_B - \rho_A)\Delta y,$$

with a similar expression for $\rho(F)$. Finally, the interpolated density at G between E and F is given by

$$\rho(G) \cong \rho_G = \rho_E + (\rho_F - \rho_E)\Delta z.$$

Putting all these together, it is easy to show that

$$\begin{aligned}\rho_G = \rho_{000} &+ \Delta x(\rho_{100} - \rho_{000}) + \Delta y(\rho_{010} - \rho_{000}) + \Delta z(\rho_{001} - \rho_{000})\\ &+ \Delta x\Delta y(\rho_{000} + \rho_{110} - \rho_{100} - \rho_{010})\\ &+ \Delta y\Delta z(\rho_{000} + \rho_{011} - \rho_{010} - \rho_{001})\\ &+ \Delta z\Delta x(\rho_{000} + \rho_{101} - \rho_{001} - \rho_{100})\\ &+ \Delta x\Delta y\Delta z(\rho_{100} + \rho_{010} + \rho_{001} + \rho_{111} - \rho_{000} - \rho_{101}\\ &- \rho_{011} - \rho_{110}).\end{aligned}$$

13.4.9. Combining different crystal forms

Frequently, a molecule crystallizes in a variety of different crystal forms [*e.g.* hexokinase (Fletterick & Steitz, 1976), the influenza virus neuraminidase spike (Varghese *et al.*, 1983), the histocompatibility antigen HLA (Bjorkman *et al.*, 1987) and the CD4 receptor (Wang *et al.*, 1990)]. It is then advantageous to average between the different crystal forms. This can be achieved by averaging each crystal form independently into a standard orientation in the h-cell (if the redundancy is $N = 1$ for a given crystal form, then this simply amounts to producing a skewed representation of the p-cell in the h-cell environment). The different results, now all in the same h-cell orientation, can be averaged. However, care must be taken to put equal weight on each molecular copy. If the ith cell contains N_i noncrystallographic copies, then the average of the densities, $\rho_i(\mathbf{x})$ $(i = 1, 2, \ldots, I)$, is

$$\sum_i N_i\rho_i(\mathbf{x}) \Big/ \sum_i N_i$$

at each grid point, $\mathbf{x}$, in the h-cell. Additional weights can be added to account for the subjective assessment of the quality of the electron densities in the different crystal cells.

With the h-cell density improved by averaging among different crystal forms, it can now be replaced into the different p-cells. These p-cells can then be back-transformed in the usual manner to obtain a better set of phases. These, in turn, can be associated with the observed structure amplitudes for each p-cell structure, and the cycle can be repeated.

13.4.10. Phase extension and refinement of the NCS parameters

Fourier back-transformation of the modified (averaged and solvent-flattened) map leads to poor phase information immediately outside the previously used resolution limit. If no density modification had been made, the Fourier transform would have yielded exactly the same structure factors as had been used for the original map. However, the modifications result in small structure amplitudes just beyond the previous resolution limit. The resultant phases can then be used in combination with the observed amplitudes in the next map calculation, thus extending the limit of resolution.

If the cell edge of an approximately cubic unit cell is a, and the approximate radius of the molecule is $\mathcal{R}$ (therefore, $\mathcal{R} < a$), then the first node of a spherical diffraction function will occur when $H\mathcal{R} = 0.7$, where H is the length of the reciprocal-lattice vector between the closest previously known structure factor and the structure factor just outside the resolution limit. Let $H = n(1/a)$, and let it be assumed that the diffraction-function amplitude is negligible when $H\mathcal{R} > 0.7$. Thus, for successful extension, $n = a/\mathcal{R}$. In general, that means that phase extension should be less than two reciprocal-lattice units in one step.

As phase extension proceeds, the accuracy of the NCS elements and the boundaries of the envelope must be constantly improved and updated to match the improved resolution. Arnold & Rossmann (1986, 1988) discussed phase error as a function of error in the NCS definition and applied rigid-body least-squares refinement for refining particle position and orientation of human rhinovirus 14. The 'climb' procedure has been found especially useful (Muckelbauer *et al.*, 1995). This depends upon searching one at a time for the parameters (rotational and translational) that minimize the near r.m.s. deviation of the individual densities to the resultant averaged densities.

Improvement of the NCS parameters is dependent upon an accurate knowledge of the cell dimensions. In the absence of such knowledge, the rotational NCS relationship cannot be accurate, since elastic distortion will result, leading to very poor averaged density. This was the case in the early determination of southern bean mosaic virus (Abad-Zapatero *et al.*, 1980), where the structure solution was probably delayed at least one year due to a lack of accurate cell dimensions.

Another aspect to phase extension is the progressive decrease in or quality of observed structure amplitudes. The observed amplitudes can be augmented with the calculated values obtained by Fourier back-transformation of the averaged map. However, clearly, as the number of calculated values increases in proportion to the number of observed values, the rate of convergence decreases. In the limit, when there are no available $F_{\rm obs}$ values, averaging a map based on $F_{\rm calc}$ values will not alter it, and, thus, convergence stops entirely.

13.4.11. Convergence

Iterations consist of averaging, Fourier inversion of the average map, recombination of observed structure-factor amplitudes with calculated phases, and recalculation of a new electron-density map. Presumably, each new map is an improvement of the previous map as a consequence of using the improved phases resulting from the map-averaging procedure. However, after five or ten cycles, the procedure has usually converged so that each new map is essentially the same as the previous map. Convergence can be usefully measured by computing the correlation coefficient (*CC*) and *R* factor (*R*) between calculated ($F_{\rm calc}$) and observed ($F_{\rm obs}$) structure-factor amplitudes as a function of resolution (Fig. 13.4.11.1). These factors are defined as

$$CC = \frac{\sum_h(\langle F_{\rm obs}\rangle - F_{\rm obs})(\langle F_{\rm calc}\rangle - F_{\rm calc})}{\left[\sum_h(\langle F_{\rm obs}\rangle - F_{\rm obs})^2(\langle F_{\rm calc}\rangle - F_{\rm calc})^2\right]^{1/2}},$$

$$R = 100 \times \sum\left|\left(|F_{\rm obs}| - |F_{\rm calc}|\right)\right| \Big/ \sum|F_{\rm obs}|.$$

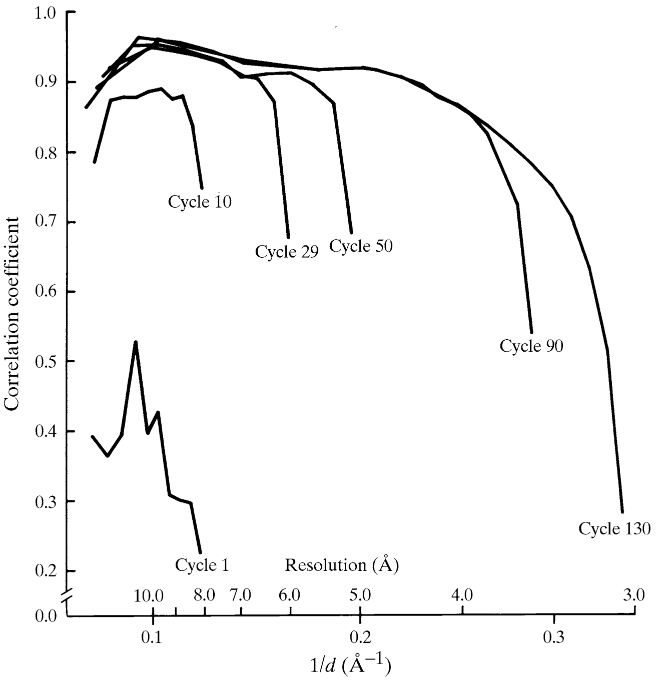


Figure 13.4.11.1

Plot of a correlation coefficient as the phases were extended from 8 to 3 Å resolution in the structure determination of Mengo virus. [Reproduced with permission from Luo *et al.* (1989).]

Because of the lack of information immediately outside the resolution limit, these factors must necessarily be poor in the outermost resolution shell. Nevertheless, the outermost resolution shell will be the most sensitive to phase improvement as these structure factors will be the furthest from their correct values at the start of a set of iterations after a resolution extension.

Convergence of *CC* and *R* does not, however, necessarily mean that phases are no longer changing from cycle to cycle. Usually, the small-amplitude structure factors keep changing long after convergence appears to have been reached (unpublished results). However, the small-amplitude structure factors make very little difference to the electron-density maps.

The rate of convergence can be improved by suitably weighting coefficients in the computation of the next electron-density map. It can be useful to reduce the weight of those structure factors where the difference between observed and calculated amplitudes is larger than the average difference, as, presumably, error in amplitude can also imply error in phase. Various weighting schemes are generally used (Sim, 1959; Rayment, 1983; Arnold *et al.*, 1987; Arnold & Rossmann, 1988).

As mentioned above, the rate of convergence can also be improved by inclusion of $F_{\rm calc}$ values when no $F_{\rm obs}$ values have been measured. However, care must be taken to use suitable weights to ensure that the $F_{\rm calc}$'s are not systematically larger or smaller than the $F_{\rm obs}$ values in the same resolution range.

Monitoring the *CC* or *R* factor for different classes of reflections (*e.g.* $h + k + l = 2n$ and $h + k + l = 2n + 1$) can be a good indicator of problems (Muckelbauer *et al.*, 1995), particularly in the presence of pseudo-symmetries. All classes of reflections should behave similarly.

The power (P) of the phase determination and, hence, the rate of convergence and error in the final phasing has been shown to be (Arnold & Rossmann, 1986) proportional to

$$P \propto (Nf)^{1/2}/[R(U/V)],$$

where N is the NCS redundancy, f is the fraction of observed reflections to those theoretically possible, R is a measure of error on the measured amplitudes (*e.g.* $R_{\rm merge}$) and U/V is the ratio of the volume of the density being averaged to the volume of the unit cell. Important implications of this relationship include that the phasing power is proportional to the square root of the NCS redundancy and that it is also dependent upon solvent content and diffraction-data quality and completeness.

13.4.12. *Ab initio* phasing starts

Some initial low-resolution model is required to initiate phasing at very low resolution. The use of cryo-EM reconstructions or available homologous structures is now quite usual. However, a phase determination using a sphere or hollow shell is also possible. In the case of a spherical virus, such an approximation is often very reasonable, as is evident when plotting the mean intensities at low resolution. These often show the anticipated distribution of a Fourier transform of a uniform sphere (Fig. 13.4.12.1). Thus, initiating phasing using a spherical model does require the prior determination of the average radius of the spherical virus. This can be done either by using an R-factor search (Tsao, Chapman & Rossmann, 1992) or by using low-angle X-ray scattering data (Chapman *et al.*, 1992). A minimal model would be to estimate the value of $F(000)$ on the same relative scale as the observed amplitudes. This structure factor must always have a positive value. Such a limited initial start was first explored by Rossmann & Blow (1963).

In surprisingly many cases (Valegård *et al.*, 1990; Chapman *et al.*, 1992; McKenna, Xia, Willingmann, Ilag, Krishnaswamy *et al.*, 1992; McKenna, Xia, Willingmann, Ilag & Rossmann, 1992; Tsao, Chapman & Rossmann, 1992; Tsao, Chapman, Wu *et al.*, 1992), it has been found that initiating phasing by using a very low resolution model results in a phase solution of the Babinet inverted structure ($\alpha \rightarrow \alpha + \pi$), where the desired density is negative instead of positive. Presumably, this is the result of phase convergence in a region where the assumed spherical transform is π out of step with reality. As long as this possibility is kept in mind with a watchful eye, such an inversion does not hamper good phase determination. In the case of phase extension, stepping too far in resolution can also lead to analogous problems (Arnold *et al.*, 1987).

Similar errors can occur due to lack of information on the correct enantiomorph in the initial phasing model. In some cases, where spherical envelopes are used and the distribution of NCS elements is also centric, there will be no decision on hand, and the phases will remain centric (Johnson *et al.*, 1975). However, in general, the enantiomorphic ambiguity (hand assignment) can be resolved by providing a model that has some asymmetry or by arbitrarily selecting the phase of a large-amplitude structure factor away from its centric value.

The progress of phase refinement away from false solutions has been the subject of 'post mortem' examinations (Valegård *et al.*, 1990; Chapman *et al.*, 1992; McKenna, Xia, Willingmann, Ilag, Krishnaswamy *et al.*, 1992; McKenna, Xia, Willingmann, Ilag & Rossmann, 1992; Tsao, Chapman & Rossmann, 1992; Tsao, Chapman, Wu *et al.*, 1992; Dokland *et al.*, 1998). The main lesson learned from these observations is that phase determination using NCS is amazingly powerful. Most initial errors in phasing gradually work themselves out with subsequent iterations and phase extension.

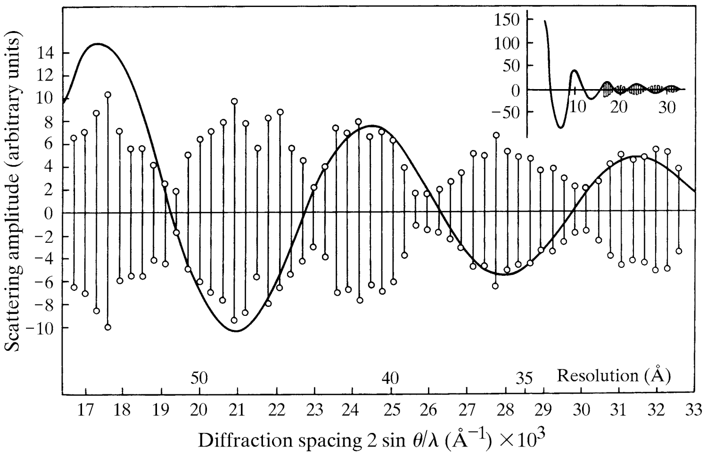


Figure 13.4.12.1
Structure amplitudes of the type II crystals of southern bean mosaic virus, averaged within shells of reciprocal space, shown in relation to the Fourier transform of a 284 Å diameter sphere. The inset shows the complete spherical transform from infinity to 30 Å resolution. [Reproduced with permission from Johnson *et al.* (1976). Copyright (1976) Academic Press.]

Perhaps the power of NCS phase determination should not be overly surprising. When phases are determined by multiple isomorphous replacement, the amount of data collected for the given molecular weight is $(N+1)$, where N is the number of derivatives and is usually 3 or 4. Similarly, for multiwavelength anomalous-dispersion data collection, there might be measurements at four different wavelengths, essentially giving $N = 8$ data points for each reflection. However, icosahedral virus determination frequently provides $N = 60$ data points for the equivalent resolution.

13.4.13. Recent salient examples in low-symmetry cases: multidomain averaging and systematic applications of multiple-crystal-form averaging

When averaging molecules that have segmental flexibility, it is essential to be able to define the extents of and noncrystallographic relationships among multiple segments which can flexibly reorient. No general protocol has been described for determining the minimum size or optimal number of segments to use in such cases. If the number of segments used for averaging is too small, then the NCS parameters cannot accurately superpose the entirety of the related segments. If too many segments are used for averaging, the segments may become too small for accurate determination of the NCS parameters. The use of too many segments may also become awkward and somewhat inefficient, since in some program systems the total number of maps that must be stored in a given cycle of averaging is proportional to the number of segments used for averaging. Comparison of atomic models for related segments that have been built or refined independently may provide convenient definitions of envelopes for averaging. In practice, a radius of 2 Å or more (depending upon the stage of structure solution and completeness and expected reliability of the model) may be added around the atoms used to define a molecular mask or envelope used in averaging. As with other averaging procedures, multidomain and multiple-crystal-form averaging approaches generally benefit from updating the molecular masks as structure determination progresses.

Often, a macromolecule can be crystallized in multiple crystal forms. Advances in crystallization technology leading to the frequent occurrence of multiple crystal forms, coupled with the availability of convenient programs, have led to increasing frequency of application of multiple-crystal-form averaging for structure solution.

Proteins, especially those containing more than one folded domain, often contain flexible hinges. As long as the boundaries of and noncrystallographic relationships among the related domains in multiple copies can be determined, then density averaging can be used to improve phasing. Programs such as *O* can be conveniently used to obtain the initial transformations necessary for correct superposition of related segments. NCS parameters can be refined using routines that either minimize the density differences among related copies or that perform rigid-body refinements of atomic models.

A number of experimental techniques have been described that may permit more widespread application of multiple-domain and multiple-crystal-form averaging. Freezing of macromolecular crystals to liquid-nitrogen temperatures has become a routine approach for enhancing the resolution and quality of macromolecular X-ray diffraction data. With most macromolecular crystals, there is a shrinkage of the 'frozen' unit cell relative to the lattice of the 'unfrozen' crystals. In many cases, significantly different cell dimensions can also be obtained by using different cryo-protective buffer and salt conditions. These variations can be exploited in a systematic fashion for phasing by electron-density averaging, so long as (1) the shrinkage relationships among the different crystals are not merely isotropic and (2) the boundaries and NCS parameters among related segments can be determined. Perutz (Perutz, 1946; Bragg & Perutz, 1952) recognized the potential utility of such shrinkage stages for crystallographic phasing in studies of haemoglobin crystals with varying degrees of hydration.

Recent examples of structure solutions involving multidomain and multiple-crystal-form averaging include studies of HIV reverse transcriptase (RT) (Ren *et al.*, 1995; Ding *et al.*, 1995). Studies of HIV RT by Stuart and coworkers involved multidomain and multiple-crystal-form averaging using different soaking solutions (Esnouf *et al.*, 1995; Ren *et al.*, 1995), in some cases with dramatically improved diffraction resolution. Arnold and coworkers have applied multidomain and multiple-crystal-form averaging to studies of HIV RT, including a systematic application of averaging electron density between 'frozen' and 'unfrozen' crystal forms (Ding *et al.*, 1995; Das *et al.*, 1996). Tong *et al.* (1997) recently described electron-density averaging among multiple closely related crystal forms of the human cytomegalovirus protease that were obtained by treatment of the crystals with different soaking buffers containing differing levels of precipitants, such as salt and polyethylene glycol.

13.4.14. Programs

This review hopefully covers most aspects encountered when employing electron-density averaging, yet the authors have drawn liberally from their own experience. There are now a large number of averaging programs and procedures available, some more suitable for structure determinations of proteins with low NCS redundancy and improper relationships (Jones, 1992) and others particularly suitable for high NCS redundancy, such as is encountered in the study of icosahedral viruses. For large structures, phase determination can be a very time-consuming computer operation. Therefore, attempts have been made to parallelize some programs (Cornea-Hasegan *et al.*, 1995), although this may lead to difficulties in exporting the programs to new and different computers.

Recently described program packages for symmetry averaging have been successfully applied to a number of cases. General program systems for averaging that are well suited to cases with high NCS include *ENVelope* (Rossmann *et al.*, 1992) and *GAP* (Jonathan Grimes and David Stuart, unpublished results); these same packages have also been used for multiple-crystal-form averaging and problems with low symmetry. A number of the program packages have been conveniently integrated with interactive computer-graphics programs such as *O* (Jones *et al.*, 1991) and most permit molecular-envelope definition by a number of possible approaches. *RAVE* and *MAVE* (Kleywegt & Jones, 1994), programs for graphics-assisted averaging within and between crystal forms, also come with an array of tools for flexible map handling and envelope definition (Kleywegt & Jones, 1996). The program systems *DMMULTI* (Cowtan & Main, 1993) and *MAGICSQUASH* (Schuller, 1996), which both derive from the program *SQUASH* (Zhang, 1993), can simultaneously apply real-space (symmetry averaging and solvent levelling with or without histogram matching) and reciprocal-space (phase refinement by the Sayre equation) constraints for phase improvement and extension. The advantage of adding phasing by the Sayre equation is greater at higher resolution, but appears to be significant in some cases, even at relatively low resolution (Cowtan & Main, 1993). *MAGICSQUASH* has been used to determine a number of structures which required multiple-domain and multiple-crystal-form averaging (Schuller, 1996). The *DEMON/ANGEL* package allows noncrystallographic averaging among multiple crystal forms together with solvent flattening and histogram matching (Vellieux *et al.*, 1995). Other versatile programs for electron-density averaging include *AVGSYS* (Bolin *et al.*, 1993) and *PHASES* (Furey & Swaminathan, 1990, 1997), both of which have features for facilitating definition and refinement of NCS parameters.

We are most grateful to Sharon Wilder and Cheryl Towell for extensive help in creating this manuscript. We are also grateful for decades of financial support by the National Science Foundation and the National Institutes of Health during the development of the techniques reported here.

References

Abad-Zapatero, C., Abdel-Meguid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossmann, M. G., Suck, D. & Tsukihara, T. (1980). *Structure of southern bean mosaic virus at 2.8 Å resolution. Nature (London)*, **286**, 33–39.

Argos, P., Ford, G. C. & Rossmann, M. G. (1975). *An application of the molecular replacement technique in direct space to a known protein structure. Acta Cryst.* A**31**, 499–506.

Argos, P. & Rossmann, M. G. (1980). *Molecular replacement method*. In *Theory and Practice of Direct Methods in Crystallography*, edited by M. F. C. Ladd & R. A. Palmer, pp. 361–417. New York: Plenum.

Arnold, E. & Rossmann, M. G. (1986). *Effect of errors, redundancy, and solvent content in the molecular replacement procedure for the structure determination of biological macromolecules. Proc. Natl Acad. Sci. USA*, **83**, 5489–5493.

Arnold, E. & Rossmann, M. G. (1988). *The use of molecular-replacement phases for the refinement of the human rhinovirus 14 structure. Acta Cryst.* A**44**, 270–282.

Arnold, E., Vriend, G., Luo, M., Griffith, J. P., Kamer, G., Erickson, J. W., Johnson, J. E. & Rossmann, M. G. (1987). *The structure determination of a common cold virus, human rhinovirus 14. Acta Cryst.* A**43**, 346–361.

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987). *Structure of the human class I histocompatibility antigen, HLA-A2. Nature (London)*, **329**, 506–512.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. (1978). *Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. Nature (London)*, **276**, 362–368.

Bolin, J. T., Smith, J. L. & Muchmore, S. W. (1993). *Considerations in phase refinement and extension: experiments with a rapid and automatic procedure.* American Crystallographic Association Annual Meeting, May 23–28, Albuquerque, New Mexico, Vol. 21, p. 51.

Bragg, L. & Perutz, M. F. (1952). *The structure of haemoglobin. Proc. R. Soc. London Ser. A*, **213**, 425–435.

Bricogne, G. (1974). *Geometric sources of redundancy in intensity data and their use for phase determination. Acta Cryst.* A**30**, 395–405.

Bricogne, G. (1976). *Methods and programs for direct-space exploitation of geometric redundancies. Acta Cryst.* A**32**, 832–847.

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W. & Rossmann, M. G. (1974). *Structure determination of crystalline lobster* D*-glyceraldehyde-3-phosphate dehydrogenase. J. Mol. Biol.* **82**, 563–585.

Champness, J. N., Bloomer, A. C., Bricogne, G., Butler, P. J. G. & Klug, A. (1976). *The structure of the protein disk of tobacco mosaic virus at 5 Å resolution. Nature (London)*, **259**, 20–24.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Ab initio phase determination for spherical viruses: parameter determination for spherical-shell models. Acta Cryst.* A**48**, 301–312.

Cornea-Hasegan, M. A., Zhang, Z., Lynch, R. E., Marinescu, D. C., Hadfield, A., Muckelbauer, J. K., Munshi, S., Tong, L. & Rossmann, M. G. (1995). *Phase refinement and extension by means of non-crystallographic symmetry averaging using parallel computers. Acta Cryst.* D**51**, 749–759.

Cowtan, K. D. & Main, P. (1993). *Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. Acta Cryst.* D**49**, 148–157.

Crowther, R. A. (1969). *The use of non-crystallographic symmetry for phase determination. Acta Cryst.* B**25**, 2571–2580.

Das, K., Ding, J., Hsiou, Y., Clark, A. D. Jr, Moereels, H., Koymans, L., Andries, K., Pauwels, R., Janssen, P. A. J., Boyer, P. L., Clark, P., Smith, R. H. Jr, Kroeger Smith, M. B., Michejda, C. J., Hughes, S. H. & Arnold, E. (1996). *Crystal structure of 8-Cl and 9-Cl TIBO complexed with wild-type HIV-1 RT and 8-Cl TIBO complexed with the Tyr181Cys HIV-1 RT drug-resistant mutant. J. Mol. Biol.* **264**, 1085–1100.

Ding, J., Das, K., Tantillo, C., Zhang, W., Clark, A. D. Jr, Jessen, S., Lu, X., Hsiou, Y., Jacobo-Molina, A., Andries, K., Pauwels, R., Moereels, H., Koymans, L., Janssen, P. A. J., Smith, R. H. Jr, Kroeger Koepke, M., Michejda, C. J., Hughes, S. H. & Arnold, E. (1995). *Structure of HIV-1 reverse transcriptase in a complex with the non-nucleoside inhibitor α-APA R 95845 at 2.8 Å resolution. Structure*, **3**, 365–379.

Dodson, E. J., Gover, S. & Wolf, W. (1992). Editors. *Proceedings of the CCP4 Study Weekend. Molecular Replacement.* Warrington: Daresbury Laboratory.

Dokland, T., McKenna, R., Sherman, D. M., Bowman, B. R., Bean, W. F. & Rossmann, M. G. (1998). *Structure determination of the φX174 closed procapsid. Acta Cryst.* D**54**, 878–890.

Esnouf, R., Ren, J., Ross, C., Jones, Y., Stammers, D. & Stuart, D. (1995). *Mechanism of inhibition of HIV-1 reverse transcriptase by non-nucleoside inhibitors. Nature Struct. Biol.* **2**, 303–308.

Fletterick, R. J. & Steitz, T. A. (1976). *The combination of independent phase information obtained from separate protein structure determinations of yeast hexokinase. Acta Cryst.* A**32**, 125–132.

Furey, W. & Swaminathan, S. (1990). *PHASES: a program package for the processing and analysis of diffraction data from macromolecules.* Am. Crystallogr. Assoc. Meeting Abstracts, **18**, PA33, p. 73.

Furey, W. & Swaminathan, S. (1997). *PHASES-95: a program package for the processing and analysis of diffraction data from macromolecules. Methods Enzymol.* **277**, 590–620.

Gaykema, W. P. J., Hol, W. G. J., Vereijken, J. M., Soeter, N. M., Bak, H. J. & Beintema, J. J. (1984). *3.2 Å structure of the copper-containing, oxygen-carrying protein Panulirus interruptus haemocyanin. Nature (London)*, **309**, 23–29.

Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. (1978). *Tomato bushy stunt virus at 2.9 Å resolution. Nature (London)*, **276**, 368–373.

Hogle, J. M., Chow, M. & Filman, D. J. (1985). *Three-dimensional structure of poliovirus at 2.9 Å resolution. Science*, **229**, 1358–1365.

Johnson, J. E. (1978). *Appendix II. Averaging of electron density maps. Acta Cryst.* B**34**, 576–577.

Johnson, J. E., Akimoto, T., Suck, D., Rayment, I. & Rossmann, M. G. (1976). *The structure of southern bean mosaic virus at 22.5 Å resolution. Virology*, **75**, 394–400.

Johnson, J. E., Argos, P. & Rossmann, M. G. (1975). *Rotation function studies of southern bean mosaic virus at 22 Å resolution. Acta Cryst.* B**31**, 2577–2583.

Jones, T. A. (1992). *a, yaap, asap, @#*? A set of averaging programs.* In *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by E. Dodson, S. Gover & W. Wolf, pp. 91–105. Warrington: Daresbury Laboratory.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Kleywegt, G. J. & Jones, T. A. (1994). *Halloween, masks and bones.* In *From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. Waller, pp. 59–66. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1996). *xdlMAPMAN and xdlDATAMAN – programs for reformatting, analysis and manipulation of biomacromolecular electron-density maps and reflection data sets. Acta Cryst.* D**52**, 826–828.

Lin, Z., Konno, M., Abad-Zapatero, C., Wierenga, R., Murthy, M. R. N., Ray, W. J. Jr & Rossmann, M. G. (1986). *The structure of rabbit muscle phosphoglucomutase at intermediate resolution. J. Biol. Chem.* **261**, 264–274.

Luo, M., Vriend, G., Kamer, G. & Rossmann, M. G. (1989). *Structure determination of Mengo virus. Acta Cryst.* B**45**, 85–92.

McKenna, R., Xia, D., Willingmann, P., Ilag, L. L., Krishnaswamy, S., Rossmann, M. G., Olson, N. H., Baker, T. S. & Incardona, N. L. (1992). *Atomic structure of single-stranded DNA bacteriophage φX174 and its functional implications. Nature (London)*, **355**, 137–143.

McKenna, R., Xia, D., Willingmann, P., Ilag, L. L. & Rossmann, M. G. (1992). *Structure determination of the bacteriophage φX174. Acta Cryst.* B**48**, 499–511.

Main, P. (1967). *Phase determination using non-crystallographic symmetry. Acta Cryst.* **23**, 50–54.

Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Three-dimensional structure of tosyl-α-chymotrypsin. Nature (London)*, **214**, 652–656.

Muckelbauer, J. K., Kremer, M., Minor, I., Tong, L., Zlotnick, A., Johnson, J. E. & Rossmann, M. G. (1995). *Structure determination of coxsackievirus B3 to 3.5 Å resolution. Acta Cryst.* D**51**, 871–887.

Muirhead, H., Cox, J. M., Mazzarella, L. & Perutz, M. F. (1967). *Structure and function of haemoglobin. III. A three-dimensional Fourier synthesis of human deoxyhaemoglobin at 5.5 Å resolution. J. Mol. Biol.* **28**, 117–156.

Nordman, C. E. (1980). *Procedures for detection and idealization of non-crystallographic symmetry with application to phase refinement of the satellite tobacco necrosis virus structure. Acta Cryst.* A**36**, 747–754.

Perutz, M. F. (1946). *Trans. Faraday Soc.* **42B**, 187.

Rayment, I. (1983). *Molecular replacement method at low resolution: optimum strategy and intrinsic limitations as determined by calculations on icosahedral virus models. Acta Cryst.* A**39**, 102–116.

Ren, J., Esnouf, R., Garman, E., Somers, D., Ross, C., Kirby, I., Keeling, J., Darby, G., Jones, Y., Stuart, D. & Stammers, D. (1995). *High resolution structures of HIV-1 RT from four RT-inhibitor complexes. Nature Struct. Biol.* **2**, 293–302.

Rossmann, M. G. (1972). Editor. *The Molecular Replacement Method.* New York: Gordon & Breach.

Rossmann, M. G. (1990). *The molecular replacement method. Acta Cryst.* A**46**, 73–82.

Rossmann, M. G. & Arnold, E. (2001). *Patterson and molecular-replacement techniques.* In *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.3. Dordrecht: Kluwer Academic Publishers.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., Rueckert, R. R., Sherry, B. & Vriend, G. (1985). *Structure of a* *human common cold virus and functional relationship to other picornaviruses. Nature (London)*, **317**, 145–153.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Rossmann, M. G. & Blow, D. M. (1963). *Determination of phases by the conditions of non-crystallographic symmetry. Acta Cryst.* **16**, 39–45.

Rossmann, M. G., Blow, D. M., Harding, M. M. & Coller, E. (1964). *The relative positions of independent molecules within the same asymmetric unit. Acta Cryst.* **17**, 338–342.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *Molecular replacement real-space averaging. J. Appl. Cryst.* **25**, 166–180.

Schuller, D. J. (1996). *MAGICSQUASH: more versatile non-crystallographic averaging with multiple constraints. Acta Cryst.* D**52**, 425–434.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst.* **12**, 813–815.

Tong, L., Qian, C., Davidson, W., Massariol, M.-J., Bonneau, P. R., Cordingley, M. G. & Lagacé, L. (1997). *Experiences from the structure determination of human cytomegalovirus protease. Acta Cryst.* D**53**, 682–690.

Tong, L. & Rossmann, M. G. (1995). *Reciprocal-space molecular-replacement averaging. Acta Cryst.* D**51**, 347–353.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Ab initio phase determination for viruses with high symmetry: a feasibility study. Acta Cryst.* A**48**, 293–301.

Tsao, J., Chapman, M. S., Wu, H., Agbandje, M., Keller, W. & Rossmann, M. G. (1992). *Structure determination of monoclinic canine parvovirus. Acta Cryst.* B**48**, 75–88.

Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1990). *The three-dimensional structure of the bacterial virus MS2. Nature (London)*, **345**, 36–41.

Varghese, J. N., Laver, W. G. & Colman, P. M. (1983). *Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. Nature (London)*, **303**, 35–40.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification. J. Appl. Cryst.* **28**, 347–351.

Wang, J., Yan, Y., Garrett, T. P. J., Liu, J., Rodgers, D. W., Garlick, R. L., Tarr, G. E., Husain, Y., Reinherz, E. L. & Harrison, S. C. (1990). *Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. Nature (London)*, **348**, 411–418.

Wilson, I. A., Skehel, J. J. & Wiley, D. C. (1981). *Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature (London)*, **289**, 366–373.

Zhang, K. Y. J. (1993). *SQUASH – combining constraints for macromolecular phase refinement and extension. Acta Cryst.* D**49**, 213–222.

International Tables for Crystallography (2012). Vol. F, Chapter 13.5, pp. 364–366.

Chapter 13.5. Molecular replacement with *MOLREP*

A. Vagin and A. Teplyakov

13.5.1. Introduction

Molecular replacement (MR) is one of the two principal methods of crystal structure determination (Rossmann, 1972). About two-thirds of the X-ray structures recently deposited in the Protein Data Bank (Berman *et al.*, 2000) have been determined by MR. With the rapid growth in the number of protein structures available as search models, the use of the method will increase even further in the future. Traditionally, MR has been implemented as a three-dimensional rotational search followed by a three-dimensional translational search. With recent advances in computing a combined six-dimensional search is becoming feasible; however, to date the most popular program packages for MR use various algorithms for a (3 + 3)-dimensional search. These include *AMoRe* (Navaza, 2001), *Phaser* (McCoy *et al.*, 2007) and the MR implementation in *CNS* (Brünger *et al.*, 1998; Chapter 18.6). Together with *MOLREP*, they cover over 95% of the structures solved by MR.

MOLREP is an automated program for molecular replacement that utilizes a number of original approaches to rotational and translational search and data preparation. *MOLREP* was initially developed (Vagin, 1989) as a fast and efficient alternative to the few existing program packages and has subsequently been enriched with new algorithms and adapted to various operating systems. *MOLREP* was originally part of the program suite *BLANC* (Vagin, 1982; Vagin *et al.*, 1998) and was later included in the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). *MOLREP* is a component of several MR pipelines, including *BALBES* (Long *et al.*, 2008), *MrBUMP* (Keegan & Winn, 2008) and the JCSG pipeline (Schwarzenbacher *et al.*, 2008), that utilize the entire PDB for fully automatic structure determination. Since the first publication describing the program (Vagin & Teplyakov, 1997) it has been cited over 3000 times. In this overview, we highlight unique features of the program, which include weighting of the X-ray data and search models, multi-copy search, fitting the model into electron density and structural superposition of two models by using the spherically averaged and phased translation function and rotation function.

13.5.2. Program operation

The philosophy behind *MOLREP* was to provide a tool such that most if not all of the tasks in MR could be performed automatically, or at least the optimal choice of parameters at each step of the process would be performed by the program. To make the program user-friendly, the interface was organized so that the user can run the program in a dialogue mode. The prompts appear as self-explanatory questions with a choice of possible answers and a list of keywords to define certain parameters. The batch mode using a command file or a command string is also available, as well as the *CCP4i* GUI.

MOLREP can perform a variety of tasks that require rotational and/or positional search: standard MR, multi-copy search, fitting a model into electron density, heavy-atom search and model superposition. The arsenal of rotation (RF) and translation (TF) functions includes self-RF, cross-RF, locked cross-RF, phased RF, full-symmetry TF, phased TF, spherically averaged phased TF and packing function (PF). The program is general for all space groups.

MR with *MOLREP* is not only fast and efficient but is also largely automatic. The minimum input provided by the user includes the X-ray data and the atomic model.

The input reflection file may be in *CCP*4 MTZ format, *BLANC* format or CIF format. An electron-density map or an electron-microscopy (EM) reconstruction (in *CCP*4 or *BLANC* format) may substitute for the structure factors. In this case, the map is converted to structure factors and phases. Crystal symmetry information is taken from the reflection file.

The input model is usually provided as an X-ray structure in the PDB format, but may also be an NMR ensemble. The ensemble is utilized either as a single entity (structure factors are calculated from the entire ensemble) or as individual members (MR is carried out for each structure). A set of homologous structures can be used in a similar way to the NMR ensemble. An electron-microscopy image or electron-density map can also substitute for the search model.

MOLREP can run with a minimum input using default parameters and parameters calculated from the input data. However, it may be useful to control the program through the use of a wide range of keywords to override the defaults. The keywords are described in the documentation supplied with the software.

The output of the program is a PDB file with the atomic model ready for refinement and a text file with details of the calculations. Additionally, an XML file suitable for communication between different programs in a pipeline is generated.

13.5.3. Preparation of the search model

The search model is automatically modified in various ways in order to try these variants for MR (Lebedev *et al.*, 2008). The options include (i) a polyalanine model, (ii) a model with modified atomic B factors increased proportionally to the atom accessibility and (iii) a model with the corrected amino-acid sequence according to the alignment of the search and target sequences. The alignment is carried out by the program when the target sequence is supplied. It takes into account the three-dimensional structure of the search model and weights the buried residues more than the surface residues. It also positions insertions and deletions away from the secondary-structure elements. Based on this alignment, certain atoms in the side chains and entire residues in the deletion regions are removed from the search model.

13.5.4. Preparation of the X-ray data

Proper treatment of the X-ray data may dramatically improve the results of MR and structure refinement. Several unique and conventional approaches for data scaling and correction have been implemented in *MOLREP*.

Anisotropic correction of the experimental data is applied when the X-ray intensity fall-off with resolution varies substantially with direction. The X-ray data are fitted to the isotropic Gaussian derived from the origin peak of the Patterson function (Rogers, 1965). The procedure is implemented in reciprocal space according to Blessing & Langs (1988).

Scaling of the observed and calculated structure factors is based on the scaling of the height and width of the corresponding Patterson origin peaks (Rogers, 1965; Blessing & Langs, 1988). This method is advantageous when only low-resolution data are available. In such a case, the estimation of the overall B factor ($B_{\rm over}$) from the Wilson plot may be inaccurate. Scaling by Patterson is also of value for the cross-RF, where different cells are used for the search model and for the unknown structure.

To define the weighting scheme for the X-ray data, *MOLREP* estimates a number of parameters for the search model. Two of them, the radius of gyration of the model (R_g) and the sequence similarity to the target protein (Ω), are translated into the parameters of a low-pass filter and a high-pass filter (Gonzalez & Woods, 2002), which define the resolution dependence of the reflection weights in the RF and TF. It is assumed that the intensities of the low-resolution reflections mostly depend on large-scale details of the crystal and are not very sensitive to the internal features of the molecules. In contrast, the high-resolution reflections define the structural details that are smaller than the differences between the search and the target molecule. The exact measure of this difference is unknown until the structure is solved, but it correlates with the known sequence similarity (Chothia, 1992). The low-pass filter helps to reduce noise in the RF and TF by dampening high-resolution reflections and thus blurring the Patterson map. The low-pass filter is applied to structure factors as an additional B factor,

$$F_{\rm new} = F \exp[-B_{\rm add}(\sin\theta/\lambda)^2]. \qquad (13.5.4.1)$$

The value of $B_{\rm add}$ is based on the sequence similarity and normalized to the Patterson origin peak according to an empirical formula,

$$B_{\rm add} = (B_{\rm lim} - B_{\rm res})(1-\Omega)^4 + B_{\rm res} - B_{\rm over}, \qquad (13.5.4.2)$$

where $B_{\rm lim}$ is the maximum allowed addition to the B factor when no similarity exists ($\Omega = 0$). $B_{\rm lim} = 8\pi^2 u^2$, where u is the mean atomic displacement (1.1 Å). $B_{\rm res}$ is determined from the maximum resolution $R_{\rm max}$ as

$$B_{\rm res} = 2R^2_{\rm max}. \qquad (13.5.4.3)$$

The high-pass filter dampens low-resolution reflections and thus sharpens the Patterson map. The high-pass filter is applied to structure factors as

$$F_{\rm new} = F\{1 - \exp[-B_{\rm off}(\sin\theta/\lambda)^2]\}, \qquad (13.5.4.4)$$

where $B_{\rm off} = 2\pi^2(R_g/6)^2$.

The filter parameters derived automatically from the search model proved to work well in some difficult MR test cases. However, the user may want to define them explicitly when, for example, high sequence similarity between the search and the target protein is in discord with the significant differences in their structures.

13.5.5. Rotational search

The rotational search is performed using the RF of Crowther (1972), which utilizes the fast Fourier transform (FFT) technique. The number of spherical harmonics used in the calculations is limited to 100. Spherical harmonics with $l = 0$ are not included in order to eliminate the effect of the origin Patterson peak. The default radius of the integration sphere is derived from the size of the search model and is usually two times larger than the radius of gyration. However, the Patterson radius should not significantly exceed half of the minimal unit-cell dimension because the contribution of intermolecular vectors becomes significant.

The RF solutions may be refined prior to positional search using a rigid-body technique. The refinement is performed in space group $P1$ and the outcome is evaluated by the correlation coefficient.

13.5.6. Positional search

The full-symmetry TF (Vagin, 1989) originates from the T2 function of Crowther & Blow (1967) corrected by Harada *et al.* (1981). It simultaneously uses all symmetry operators, resulting in a single peak with an improved signal-to-noise ratio which directly gives the position of the model in the unit cell. In addition, the TF is coupled with a PF to remove false maxima which correspond to interpenetrating molecules. Both the TF and PF allow the incorporation of a second model already placed in the cell. The TF solution may be subjected to rigid-body refinement incorporated in *MOLREP*. Noncrystallographic symmetry may be imposed on the model in order to restrain the refinement.

Pseudo-translation is automatically detected from analysis of the Patterson map. A significant off-origin peak gives the pseudo-translation vector, which is used to modify structure factors in the TF calculation (Navaza *et al.*, 1998). Alternatively, the pseudo-translation vector may be supplied by the user. A pseudo-translational copy of the model is added at the end of the positional search.

13.5.7. Multi-copy search

The MR method has been extended to a simultaneous search for multiple copies of the macromolecule in the unit cell (Vagin & Teplyakov, 2000). The central point of this approach is the construction of a dimer search model from the properly oriented monomers using a special TF. This model is then used for a positional search with a conventional TF. The method does not impose any limitation on the oligomeric structure of the protein, either on the number of monomers or on their relative location, *i.e.* pure rotational symmetry is not required. In principle, the monomers may even be of different types. In the case of two monomers, which we call a dyad, the multi-copy search takes place as follows.

(i) The RF is calculated for a monomeric search model. N_R highest peaks of the RF are used to produce a set of oriented monomers. All possible $N_R(N_R + 1)/2$ pair combinations of these monomers constitute a set of putative dyads.

(ii) For each putative dyad, the intermolecular vector relating the two monomers of the dyad is determined using a special TF. The special TF is a phased TF, which treats the Patterson function as electron density, and the search model is described by the structure factors $F_1F_2^*$. The solution of the special TF is the dyad vector. N_T top solutions will be considered to ensure that the correct dyad is not missed. Therefore, the total number of dyads selected for the final positional search will be $N_TN_R(N_R + 1)/2$.

(iii) A positional search for each dyad is performed using a conventional TF. The results are estimated on the basis of the TF value and a correlation coefficient.

The dyad search can be extended to a triad search by including a third monomer in the search model. The phased TF will then be used to find three vectors describing the triad. However, the dyad search seems to be sufficient in most cases, as the main problem is usually the location of the first pair of monomers. When this task is fulfilled, the search can be repeated for a third monomer or a second dyad with the first being fixed.

Although there are no limitations on the relative position and orientation of the two monomers constituting a dyad, the search space can be limited by imposing restraints on the oligomeric structure, *e.g.* by defining the pure rotational symmetry. This feature may be particularly useful in the case of molecular dimers and higher oligomers. Prior knowledge of the molecular symmetry may not only reduce the computational time but may also facilitate the search by selecting functionally meaningful solutions. This can be achieved by defining the angular relation between the monomers, *e.g.* as a $180 \pm 10^\circ$ rotation. Alternatively, the search may be reduced to the use of several top peaks of the self-RF that define a set of possible relative orientations of the monomers in a dyad.

13.5.8. Fitting the model into electron density

This 'real-space' search is employed when some phase information is available either from experimental data (multiple isomorphous replacement or multiwavelength anomalous diffraction) or from a partial MR solution, *e.g.* when one domain of a multi-domain protein has been located in the unit cell. Positioning a model into electron density (or into an EM reconstruction) is carried out in three steps.

(i) The model is positioned in the cell using the spherically averaged TF (Vagin & Isupov, 2001). The locally spherically averaged experimental electron density is compared with that calculated from the model and all possible positions are tabulated.
(ii) For each selected position, the orientation of the model is determined using the local phased RF.
(iii) The orientation and position of the model is verified and refined using the phased TF.

It should be noted that a full six-dimensional search, although time consuming, may in many cases work better. However, compared with traditional (3 + 3)-dimensional searches the fitting based on the spherically averaged phased TF may be advantageous, particularly when the search model is small. In general, the procedure works better for globular models.

The same procedure is used for the superposition of atomic models. One model is converted to the electron-density map and another model is fitted to this map using the spherically averaged and phased TF and RF. No sequence information or secondary-structure assignment is required for such a superposition. The procedure may be useful for proteins with very low sequence similarity. It is also suitable for nonprotein models such as nucleic acids or carbohydrates.

13.5.9. Distribution

The program *MOLREP* is written in standard Fortran 90 and can be run under Linux, Unix, Mac OSX and Windows. It is available free to academic users as a standalone version or as part of the *CCP*4 suite. The program, examples and documentation may be downloaded from the author's web site (http://www.ysbl.york.ac.uk/~alexei/molrep.html) or from the CCP4 web site. All inquiries about the program should be addressed to Alexei Vagin (alexei@ysbl.york.ac.uk).

We thank Andrey Lebedev for his contribution to the program and helpful discussions. This work was supported by CCP4, the Biotechnology and Biological Sciences Research Council (BBSRC) and EU Biotech.

This chapter is based on a recent article published in *Acta Crystallographica Section D* (Vagin & Teplyakov, 2010).

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Blessing, R. H. & Langs, D. A. (1988). *A priori estimation of scale and overall anisotropic temperature factors from the Patterson origin peak*. *Acta Cryst.* A**44**, 729–735.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR System: a new software suite for macromolecular structure determination*. *Acta Cryst.* D**54**, 905–921.

Chothia, C. (1992). *One thousand families for the molecular biologist*. *Nature (London)*, **357**, 543–544.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography*. *Acta Cryst.* D**50**, 760–763.

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–183. New York: Gordon & Breach.

Crowther, R. A. & Blow, D. M. (1967). *A method of positioning a known molecule in an unknown crystal structure*. *Acta Cryst.* **23**, 544–548.

Gonzalez, R. C. & Woods, R. E. (2002). *Digital Image Processing.* Upper Saddle River, New Jersey: Prentice Hall.

Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *A translation function combining packing and diffraction information: an application to lysozyme (high-temperature form)*. *Acta Cryst.* A**37**, 398–406.

Keegan, R. M. & Winn, M. D. (2008). *MrBUMP: an automated pipeline for molecular replacement*. *Acta Cryst.* D**64**, 119–124.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2008). *Model preparation in MOLREP and examples of model improvement using X-ray data*. *Acta Cryst.* D**64**, 33–39.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *BALBES: a molecular-replacement pipeline*. *Acta Cryst.* D**64**, 125–132.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software*. *J. Appl. Cryst.* **40**, 658–674.

Navaza, J. (2001). *Implementation of molecular replacement in AMoRe*. *Acta Cryst.* D**57**, 1367–1372.

Navaza, J., Panepucci, E. H. & Martin, C. (1998). *On the use of strong Patterson function signals in many-body molecular replacement*. *Acta Cryst.* D**54**, 817–821.

Rogers, D. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 117–148. Oxford: Pergamon Press.

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon & Breach.

Schwarzenbacher, R., Godzik, A. & Jaroszewski, L. (2008). *The JCSG MR pipeline: optimized alignments, multiple models and parallel searches*. *Acta Cryst.* D**64**, 133–140.

Vagin, A. A. (1982). PhD Thesis. Institute of Crystallography, Moscow, Russia.

Vagin, A. A. (1989). *New translation and packing functions*. *CCP4 Newsl. Protein Crystallogr.* **29**, 117–121.

Vagin, A. A. & Isupov, M. N. (2001). *Spherically averaged phased translation function and its application to the search for molecules and fragments in electron-density maps*. *Acta Cryst.* D**57**, 1451–1456.

Vagin, A. A., Murshudov, G. N. & Strokopytov, B. V. (1998). *BLANC: the program suite for protein crystallography*. *J. Appl. Cryst.* **31**, 98–102.

Vagin, A. & Teplyakov, A. (1997). *MOLREP: an automated program for molecular replacement*. *J. Appl. Cryst.* **30**, 1022–1025.

Vagin, A. & Teplyakov, A. (2000). *An approach to multi-copy search in molecular replacement*. *Acta Cryst.* D**56**, 1622–1624.

Vagin, A. & Teplyakov, A. (2010). *Molecular replacement with MOLREP*. *Acta Cryst.* D**66**, 22–25.

International Tables for Crystallography (2012). Vol. F, Chapter 14.1, pp. 367–372.

PART 14. ANOMALOUS DISPERSION

Chapter 14.1. Heavy-atom location and phase determination with single-wavelength diffraction data

B. W. Matthews

14.1.1. Introduction

As is well known, the successful introduction of the method of isomorphous replacement by Green *et al.* (1954) was the turning point in the subsequent development of protein crystallography as we now know it.

The idea that the phases of X-ray reflections from a protein crystal could be obtained by the introduction of heavy atoms into the crystal was not new, having been suggested by J. D. Bernal in 1939 (Bernal, 1939). The isomorphous-replacement method was used as early as 1927 by Cork (1927) in studying the alums. Bokhoven *et al.* (1951) subsequently extended the method to the study of a noncentrosymmetric projection of strychnine sulfate, using what would now be termed the method of single isomorphous replacement. They also suggested that by using a double isomorphous replacement, a unique phase determination could be obtained, even for noncentrosymmetric reflections. The details of the double (or multiple) isomorphous-replacement method were worked out by Harker (1956), who introduced the very useful concept of phase circles. Another contribution which was of great practical value, and which will provide the basis for much of the subsequent discussion, is the method introduced by Blow & Crick (1959) for the treatment of errors in the isomorphous-replacement method. In addition to the determination of protein phases by the method of substitution with heavy atoms, it is now routine to supplement this information by utilizing the anomalous scattering of the substituted atoms. The underlying principles trace back to articles by Bijvoet (1954), Ramachandran & Raman (1956), and Okaya & Pepinsky (1960). The first application of the anomalous-scattering method to protein crystallography was by Blow (1958), who used the anomalous scattering of the iron atoms to determine phase information for a noncentrosymmetric projection of horse oxyhaemoglobin.

In the following discussion, we first review the classical method of phase determination by isomorphous replacement, then discuss the inclusion of single-wavelength anomalous-scattering data, and conclude by discussing the use of such data for heavy-atom location. Part of the review is based on Matthews (1970).

14.1.2. The isomorphous-replacement method

Consider a protein crystal with an isomorphous heavy-atom derivative, *i.e.* a modified crystal in which heavy atoms occupy specific sites throughout the crystal, but which is in all other respects identical to the unsubstituted 'parent' crystal. Let the structure factors of the protein crystal be $\mathbf{F}_P(\mathbf{h})$, of the isomorph be $\mathbf{F}_{PH}(\mathbf{h})$, and of the heavy atoms $\mathbf{F}_H(\mathbf{h})$. (Note: Structure *amplitudes* are indicated by italic type, *e.g.* F_P, and *vectors* by bold-face type, *e.g.* $\mathbf{F}_P$.) In practice, one can measure the structure amplitudes F_P and F_{PH}, and it is desired to obtain from these observable quantities the value of the phase angle of $\mathbf{F}_P(\mathbf{h})$ so that a Fourier synthesis showing the electron density of the protein structure may be calculated. It will be assumed, for the moment, that the positions and occupancy of the sites of heavy-atom binding have been determined as accurately as possible.

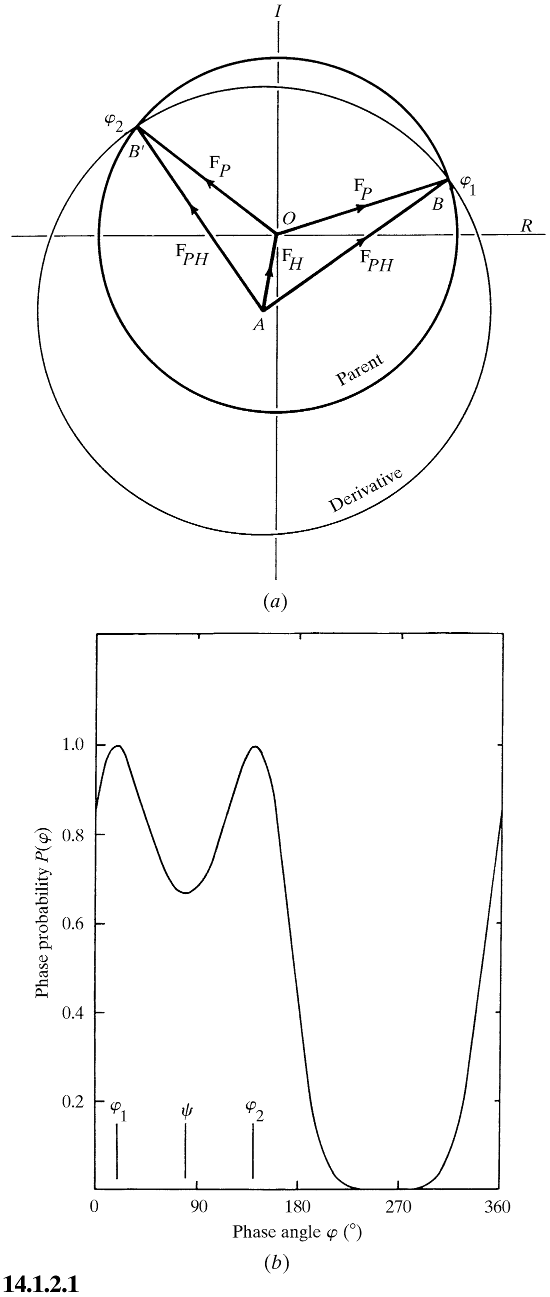


Figure 14.1.2.1
(*a*) Harker construction for a single isomorphous replacement. φ_1 and φ_2 are the 'most probable' phases for $\mathbf{F}_P$. (*b*) Phase probability distribution for a single isomorphous replacement. This and subsequent probabilities are unnormalized. [All figures in this chapter are reproduced with permission from Matthews (1970).]

Copyright © 2012 International Union of Crystallography

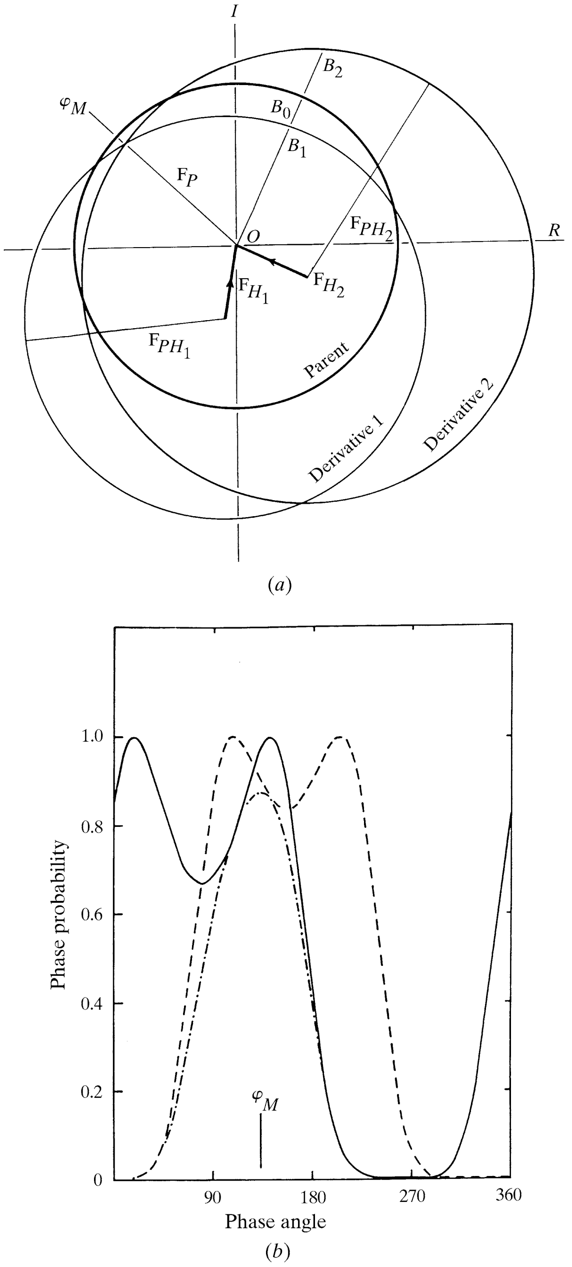


Figure 14.1.3.1
(*a*) Harker construction for a double isomorphous replacement. φ_M is the 'most probable' phase for $\mathbf{F}_P$. (*b*) Phase probability distribution corresponding to the double isomorphous replacement shown in part (*a*). The curve for derivative 1 is solid, that for derivative 2 is dashed, and that for the combined distribution is drawn as a dotted-and-dashed line.

From the heavy-atom parameters, the corresponding structure factor $\mathbf{F}_H(\mathbf{h})$ is calculated. To determine φ, the phase of $\mathbf{F}_P(\mathbf{h})$, we construct a set of phase circles, as proposed by Harker (1956). From a chosen origin O (Fig. 14.1.2.1*a*), the vector OA is drawn equal to $-\mathbf{F}_H$. Circles of radius F_P and F_{PH} are then drawn about O and A, respectively. The intersections of the phase circles at B and B' define two possible phase angles for F_P. Note that the angles are symmetrical about $\mathbf{F}_H$. This ambiguity may in principle be resolved in two ways: (*a*) by using a second heavy-atom isomorphous derivative or (*b*) by utilizing the anomalous-scattering effects for the first isomorph.

14.1.3. The method of multiple isomorphous replacement

The phase information provided by a second isomorph is illustrated in Fig. 14.1.3.1(*a*). In theory, the three phase circles will intersect at a point and the phase ambiguity will be resolved. In practice, there will be errors in the observed amplitudes F_P and F_{PH} and in the heavy-atom parameters (and thus in $\mathbf{F}_H$). Also, the isomorphism may be imperfect. As a result, the intersections of the three phase circles may not coincide. Another complication arises from the fact that for reflections where $\mathbf{F}_H$ is small, the circles will be essentially concentric and will not have well defined points of intersection. In other words, the phase determination will become indeterminate. The method of Blow & Crick (1959) was introduced as a way to take all these factors into account. It has had an extraordinary impact, not only as a practical method for protein phase determination, but also in influencing all subsequent thinking in this area.

14.1.4. The method of Blow & Crick

Blow & Crick pointed out that in practice the phase angle φ can never be determined with complete certainty. Rather, there is a finite probability that any arbitrary phase angle may be the correct one. Consider the vector diagram shown in Fig. 14.1.4.1, in which $\mathbf{F}_H$ is known and we wish to determine the probability $P(\varphi)$ that the arbitrary phase angle φ is the correct phase of $\mathbf{F}_P$. Strictly, one should allow for the possibility of errors in $\mathbf{F}_H$, F_P and F_{PH}, and should consider the probability that the vector $\mathbf{F}_P$ occupies all possible positions in the Argand diagram. However, Blow & Crick suggested that the analysis might be considerably simplified by assuming that F_P and $\mathbf{F}_H$ are known accurately and that all the error lies in the observation of F_{PH}. In other words, it was assumed that the vector $\mathbf{F}_P$ must lie on the circle of radius F_P, and the probability distribution of F_P could be evaluated as a function of φ only.

For an arbitrary phase angle φ, the phase triangle (Fig. 14.1.4.1) will not close exactly. If we define F_C to be the vector sum of $\mathbf{F}_H$ and $F_P \exp(i\varphi)$, then the lack of closure of the phase triangle is given by

$$\varepsilon = F_C - F_{PH}. \qquad (14.1.4.1)$$

Following Blow & Crick, if E is the r.m.s. error associated with the measurements, and the distribution of error is assumed to be Gaussian, then the probability $P(\varphi)$ of the phase φ being the true phase is

$$P(\varphi) = N \exp(-\varepsilon^2/2E^2), \qquad (14.1.4.2)$$

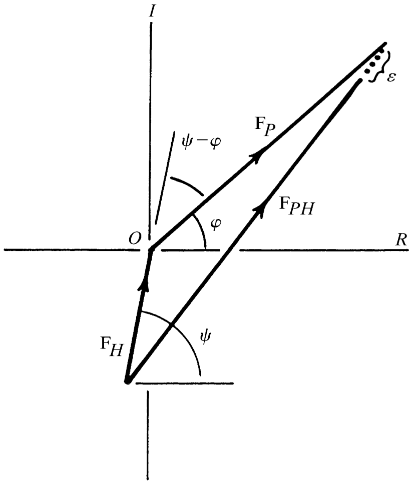


Figure 14.1.4.1
Vector diagram illustrating the lack of closure, ε, of an isomorphous-replacement phase triangle.

where N is a normalizing factor such that the sum of all probabilities is unity. The un-normalized probability distribution corresponding to Fig. 14.1.4.1 (and Fig. 14.1.2.1*a*) is shown in Fig. 14.1.2.1(*b*). The two most probable phase angles ($\varphi = \varphi_1$ and $\varphi = \varphi_2$) are the alternative phases of F_P for which the phase triangle is closed.

Individual probability distributions for the additional heavy-atom derivatives are derived in an analogous manner and may be multiplied together to give an overall probability distribution. The joint probability distribution corresponding to Fig. 14.1.3.1(*a*) is shown in Fig. 14.1.3.1(*b*), and in this case the most probable phase is that which simultaneously fits best the observed data for the two isomorphous derivatives.

The main objection which may be made to the Blow & Crick treatment is that it assumes that there is no error in F_P. In practice, however, this is not a serious limitation.

14.1.5. The best Fourier

A protein crystallographer desires to obtain a Fourier synthesis that can most readily be interpreted in terms of an atomic model of the structure. One synthesis which could be calculated is the 'most probable Fourier', obtained by choosing the value of $F_P(\mathbf{h})$ for each reflection which corresponds to the highest value of $P(\varphi)$. Blow & Crick pointed out that although this Fourier is the most likely to be correct, it has certain disadvantages. In the first place, it might tend to give too much weight to uncertain or unreliable phases, and, in the second place, for cases where $P(\varphi)$ is bimodal, there is a strong chance of making a large error in the phase angle. Blow & Crick suggested that in cases such as this, a compromise is needed, and that the centroid of the phase probability distribution provides just the required compromise. They showed that the corresponding synthesis is the 'best Fourier', which is defined to be *that Fourier transform which is expected to have the minimum mean-square difference from the Fourier transform of the true F's when averaged over the whole unit cell.*

The centroid of the phase probability distribution may be defined as a point on the phase diagram with polar coordinates (mF_P, φ_B), where φ_B is the 'best' phase angle. The quantity m, which acts as a weighting factor for F_P, is called the 'figure of merit' of the phase determination. Its magnitude, between 0 and 1, is a measure of the reliability of the phase determination.

14.1.6. Anomalous scattering

All atoms, particularly those used in preparing heavy-atom isomorphs, give rise to anomalous scattering, especially if the energy of the scattered X-rays is close to an absorption edge. The atomic scattering factor of the atom in question can be expressed as

$$\mathbf{f} = f_0 + \Delta f' + if'' = f' + if'', \qquad (14.1.6.1)$$

where f_0 is the normal scattering factor far from an absorption edge, and $\Delta f'$ and f'' are the correction terms which arise due to dispersion effects. The quantity $\Delta f'$, in phase with f_0, is usually negative, and f'', the imaginary part, is always $\pi/2$ ahead of the phase of the real part ($f_0 + \Delta f'$). It may be noted that by using different wavelengths, the term $\Delta f'$ is equivalent to a change in scattering power of the heavy atom and produces intensity differences similar to a normal isomorphous replacement, except that in this case the isomorphism is *exact* (Ramaseshan, 1964). This is the basis of the multiwavelength-anomalous-dispersion (MAD) method (Hendrickson, 1991) discussed in Chapter 14.2. Here we focus on measurements based on a single wavelength, traditionally referred to as the 'anomalous-scattering method'.

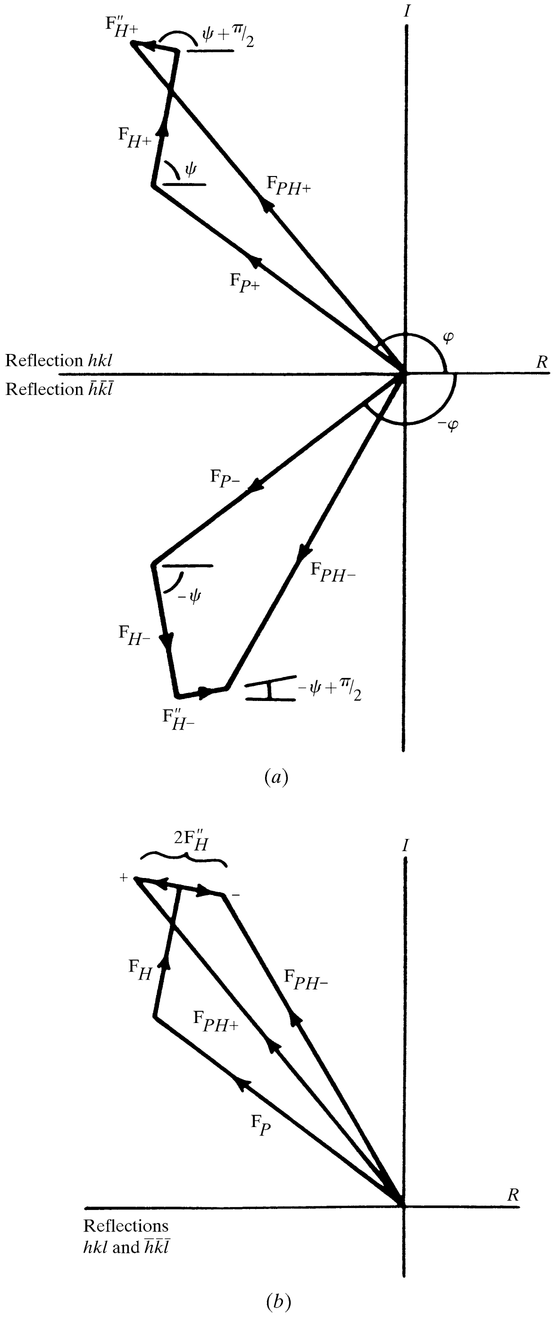


Figure 14.1.7.1
(*a*) Vector diagrams illustrating anomalous scattering for the reflections *hkl* and $\overline{hkl}$. (*b*) Combined vector diagram for reflections *hkl* and $\overline{hkl}$.

The anomalous scattering of a heavy atom is always considerably less than the normal scattering (for Cu $K\alpha$ radiation, $2f''/f'$ ranges from about 0.24 to 0.36), but there are several factors which tend to offset this reduction in magnitude (*e.g.* see Blow, 1958; North, 1965).

14.1.7. Theory of anomalous scattering

Suppose that two isomorphous crystals are differentiated by N heavy atoms of position $\mathbf{r}_n$ and scattering factor $(\mathbf{f}'_n + i\mathbf{f}''_n)$. Then, for the reflection *hkl*, the calculated structure factor of the N atoms is

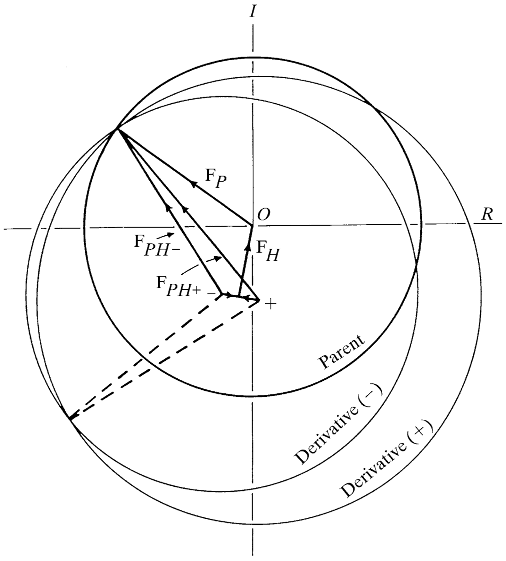


Figure 14.1.7.2
Harker construction for a single isomorphous replacement with anomalous scattering, in the absence of errors.

$$F_H(\mathbf{h}) + iF''_H(\mathbf{h}) = \sum_{n=1}^{N} f'_n(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_n) + i \sum_{n=1}^{N} f''_n(\mathbf{h}) \exp(2\pi i \mathbf{h} \cdot \mathbf{r}_n). \qquad (14.1.7.1)$$

If the heavy atoms are all of the same type, *i.e.* they all have the same ratio of f'_n/f''_n $(= k)$, then F_H and F''_H are orthogonal, and $F''_H = F_H/k$.

The relation between the structure factors of the reflection *hkl* and its Friedel mate $\bar{h}\bar{k}\bar{l}$ is illustrated in Fig. 14.1.7.1(*a*). The situation can be conveniently represented (Fig. 14.1.7.1*b*) by reflecting the $\bar{h}\bar{k}\bar{l}$ diagram through the real axis onto the *hkl* diagram. In cases such as this, where Friedel's law breaks down, we shall refer to the difference $\Delta_{PH} = (F_{PH+} - F_{PH-})$ as the Bijvoet difference, or simply the anomalous-scattering difference. The Harker phase circles corresponding to Fig. 14.1.7.1(*b*) are shown in Fig. 14.1.7.2. It will be seen that, as in the case of single isomorphous replacement, and similarly with the anomalous-scattering data alone, there is an ambiguous phase determination.

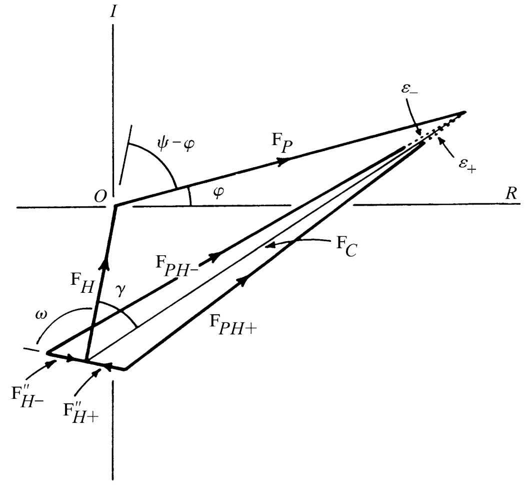


Figure 14.1.8.1
Vector diagrams illustrating lack of closure in the anomalous-scattering method.

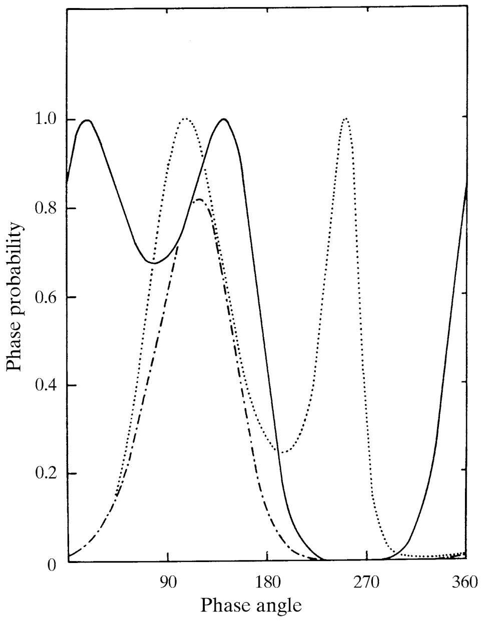


Figure 14.1.8.2
Combination of isomorphous replacement and anomalous-scattering phase probabilities for a single isomorphous replacement. $P_{\rm iso}(\varphi)$ is drawn as a solid line, $P_{\rm ano}(\varphi)$ as a dotted line, and the combined probability distribution is drawn as a dotted-and-dashed line.

In the absence of error, the three phase circles (Fig. 14.1.7.2) will meet at a point, resolving the phase ambiguity and giving a unique solution for the phase of $\mathbf{F}_P$. The isomorphous-replacement method gives phase information symmetrical about the vector $\mathbf{F}_H$, whereas the anomalous-scattering phase information for $\mathbf{F}_{PH}$ is symmetrical about $\mathbf{F}''_H$, which, for heavy atoms of the same type, is at right angles to $\mathbf{F}_H$. In other words, the two methods complement each other, one method providing exactly that information which is not given by the other.

On average, the experimentally measured isomorphous-replacement difference, $(F_{PH} - F_P)$, will be larger than the anomalous-scattering difference, $(F_{PH+} - F_{PH-})$. The former, however, relies on measurements from different crystals and is also susceptible to errors due to non-isomorphism between the parent and derivative crystals. The latter can be obtained from measurements on the same crystal, under closely similar experimental conditions, and is not affected by non-isomorphism. Therefore, it is desirable to use methods that take into account the different sources of error in the respective measurements (Blow & Rossmann, 1961; North, 1965; Matthews, 1966*b*). One method is as follows.

14.1.8. The phase probability distribution for anomalous scattering

From Fig. 14.1.8.1, it can be seen that the most probable phase angle will be the one for which $\varepsilon_+ = \varepsilon_-$. At any other phase angle, there will be an 'anomalous-scattering lack of closure' which we define to be $(\varepsilon_+ - \varepsilon_-)$. The value of $(\varepsilon_+ - \varepsilon_-)$ can readily be calculated as a function of φ (Matthews, 1966*b*; Hendrickson, 1979). Thus, if the r.m.s. error in $(\varepsilon_+ - \varepsilon_-)$ is E', and the distribution of error is assumed to be Gaussian, then from measurements of anomalous scattering, the probability $P_{\rm ano}(\varphi)$ of phase φ being the true phase of $\mathbf{F}_P$ can be estimated using an equation exactly analogous to equation (14.1.4.2).

An example of an anomalous-scattering phase probability distribution is shown by the dotted curve in Fig. 14.1.8.2. The asymmetry of the distribution arises from the fact that $P_{\rm ano}(\varphi)$ is the phase probability distribution for F_P rather than that of F_{PH}, which would be symmetrical about the phase of F''_H. The overall probability distribution obtained by combining the anomalous-scattering data with the previous isomorphous-replacement data (Fig. 14.1.2.1*b*) is given by

$$P(\varphi) = NP_{\rm iso}(\varphi)P_{\rm ano}(\varphi) \qquad (14.1.8.1)$$

and is illustrated in Fig. 14.1.8.2.

14.1.9. Anomalous scattering without isomorphous replacement

The treatment outlined above of phase determination by anomalous scattering assumed that data were available for a parent crystal devoid of anomalous scatters and an anomalously scattering isomorphous heavy-atom derivative. It is not uncommon that the native protein itself contains atoms which scatter anomalously or has been engineered to contain such scatterers. In such cases, measurements will usually be made at multiple wavelengths in order to exploit MAD phasing (Hendrickson, 1991). If, however, measurements are available only at a single wavelength, they can be utilized to obtain some phase information (*e.g.* Matthews, 1970).

14.1.10. Location of heavy-atom sites

During the development of protein crystallography, it was understood that heavy-atom sites might be located from difference Patterson functions, but there was substantial debate as to the type of function that was preferable (Perutz, 1956).

Blow (1958), and also Rossmann (1960), advocated a Patterson function with amplitudes $(F_{PH} - F_P)^2$. It relies on the admittedly crude assumption that the desired scattering amplitude of the heavy atoms, $|F_H|$, can be approximated by

$$|F_H| \simeq |F_{PH} - F_P|. \qquad (14.1.10.1)$$

The approximation does have one very helpful characteristic, namely, that it tends to be most accurate when $|F_{PH} - F_P|$ is large, *i.e.* when F_H is parallel or antiparallel to F_P (*cf.* Fig. 14.1.4.1). Thus, the numerically largest coefficients in the Patterson function tend to represent $|F_H|^2$ correctly. Given a well behaved isomorphous heavy-atom derivative, and accurately measured data, experience has shown that a map with coefficients $(F_{PH} - F_P)^2$ can give an excellent representation of the desired heavy-atom–heavy-atom vector peaks.

14.1.11. Use of anomalous-scattering data in heavy-atom location

A relation exactly analogous to equation (14.1.10.1) can be used to approximate the anomalous heavy-atom scattering amplitude, namely,

$$|F''_H| \simeq \tfrac{1}{2}|F_{PH+} - F_{PH-}| \qquad (14.1.11.1)$$

(see Fig. 14.1.7.1*b*). As noted above, if all the heavy atoms are the same, $F_H = kF''_H$. Thus, a Patterson function with coefficients $(F_{PH+} - F_{PH-})^2$ should also show the desired heavy-atom–heavy-atom vector peaks (Blow, 1957; Rossmann, 1961).

For each individual reflection, however, and as is also the case for phase determination, the information that is provided by the isomorphous-replacement difference $(|F_{PH}| - |F_P|)$ is exactly complementary to that provided by the anomalous-scattering measurement $(|F_{PH+}| - |F_{PH-}|)$. By combining both sets of experimental measurements, it is possible to obtain a much better estimate of the heavy-atom scattering, $|F_H|$, for every reflection (Kartha & Parthasarathy, 1965*a*,*b*; Matthews, 1966*a*; Singh & Ramaseshan, 1966). One formulation (Matthews, 1966*a*) can be written as

$$F_H^2 = F_P^2 + F_{PH}^2 - 2F_PF_{PH}\{1 - [wk(F_{PH+} - F_{PH-})/2F_P^2\}^{1/2}, \qquad (14.1.11.2)$$

where $F_{PH} = (F_{PH+} + F_{PH-})/2$ and w is a weighting factor (from 0 to 1) that is an estimate of the relative reliability of the measurements of $(F_{PH+} - F_{PH-})$ compared with $(F_{PH} - F_P)$.

14.1.12. Use of difference Fourier syntheses

The discussion above has focused on the use of difference Patterson functions to locate heavy-atom sites. Once one or more putative sites have been located, they can be used to calculate approximate protein phases, which, in turn, can be used to calculate difference Fourier series with coefficients in the form

$$m(F_{PH} - F_H)\exp(-i\varphi_B), \qquad (14.1.12.1)$$

where m is the figure of merit and φ_B is the 'best', albeit approximate, phase of the protein structure factor. Putting aside errors due to inaccuracies in φ_B, such maps do not give the true heavy-atom vector, $\mathbf{F}_H$. Rather, they give, essentially, the projection of $\mathbf{F}_H$ along $\mathbf{F}_P$ (*cf.* Fig. 14.1.4.1). Nevertheless, subject to certain limitations, such difference maps are extraordinarily powerful in locating secondary sites in a given heavy-atom derivative, or in using approximate phases from one derivative to search for heavy-atom sites in other putative derivatives. It is in this context, however, that certain limitations of the single-isomorphous-replacement (SIR) method have to be kept in mind. These are noted in the next section.

14.1.13. Single isomorphous replacement

Although phase determination from a single heavy-atom derivative in the absence of anomalous-scattering data is, in principle, ambiguous, it was realized early on that useful phase information can still be obtained (Blow & Rossmann, 1961). As shown in Fig. 14.1.2.1(*a*), the two possible phases for the protein are φ_1 or φ_2. In terms of the analysis of Blow & Crick (1959), the 'best' phase to use for the protein is the average of φ_1 and φ_2. This is also equivalent to using *both* φ_1 *and* φ_2. With this in mind, a situation that is of special concern is one in which the heavy-atom distribution used to determine the phases happens to have a centre of symmetry. One common way in which this can occur is when one has a heavy-atom derivative with a single site in space group $P2_1$. A related situation occurs when there are multiple sites in space group $P2_1$, but all have the same y coordinate. If the origin of coordinates is considered to be at the site of centrosymmetry, then all of the heavy-atom vectors $\mathbf{F}_H$ (Fig. 14.1.2.1*a*) will necessarily have phases of 0 or π. If such phases are used, for example, to try to identify heavy-atom-binding sites in a second derivative, the map will show the correct sites, but will also show spurious peaks of equal height related by the centre of symmetry. Faced with this choice, one must arbitrarily choose one of the alternative peaks which, in turn, will define an overall handedness for the heavy-atom arrangement. In the absence of any anomalous-scattering data, one can proceed with the structure determination in the standard way, but it must be kept in mind that either the correct electron-density map or its mirror image will ultimately be obtained.

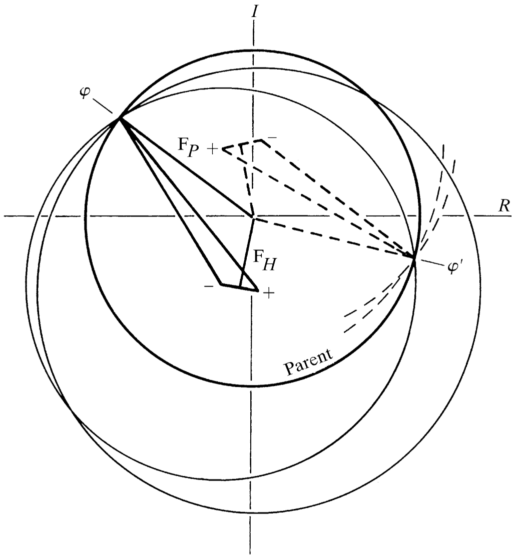


Figure 14.1.13.1

The consequences of using the incorrect hand for the heavy-atom arrangement in phase determination. The correct heavy-atom arrangement gives the construction drawn in solid lines with φ uniquely determined as the correct phase for $\mathbf{F}_P$. If the enantiomeric arrangement of heavy atoms is used, the dashed construction will result, leading to the incorrect phase φ'. Even though φ is correct and φ' is incorrect, both phase determinations have an identical figure of merit.

An alternative approach is to include anomalous-scattering data in the initial phase determination, *i.e.* to use single isomorphous replacement with anomalous scattering (SIRAS). It must be remembered that in calculating phases from anomalous-scattering data, it is first necessary to determine the coordinates of the heavy atoms *in their absolute configuration.* If the wrong hand is used in the SIRAS method (illustrated in Fig. 14.1.13.1), the resultant electron-density map will generally bear no relation to the correct electron density.

The recommended procedure, therefore, is as follows. One arbitrarily chooses one possible heavy-atom arrangement for heavy-atom derivative 1, calculates SIRAS phases and calculates a difference-electron-density map for derivative 2. The handedness of the derivative 1 coordinates are then inverted and the overall calculation repeated. The calculation based on the correct heavy-atom arrangement should show peaks at the heavy-atom sites of the second derivative. The calculation based on the incorrect arrangement shows noise (Matthews, 1966*a*). This procedure determines the absolute configuration of the heavy-atom arrangement and, at the same time, shows the derived sites for the second and subsequent derivatives.

This work was supported in part by NIH grant GM21967.

References

Bernal, J. D. (1939). *Structure of proteins. Nature (London)*, **143**, 663–667.

Bijvoet, J. M. (1954). *Structure of optically active compounds in the solid state. Nature (London)*, **173**, 888–891.

Blow, D. M. (1957). *X-ray Analysis of Haemoglobin: Determination of Phase Angles by Isomorphous Substitution*. PhD thesis, University of Cambridge.

Blow, D. M. (1958). *The structure of haemoglobin. VII. Determination of phase angles in the non-centrosymmetric [100] zone. Proc. R. Soc. London Ser. A*, **247**, 302–336.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method. Acta Cryst.* **14**, 1195–1202.

Bokhoven, C., Schoone, J. C. & Bijvoet, J. M. (1951). *The Fourier synthesis of the crystal structure of strychnine sulphate pentahydrate. Acta Cryst.* **4**, 275–280.

Cork, J. M. (1927). *The crystal structure of some of the alums. Philos. Mag.* **4**, 688–698.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. Acta Cryst.* **9**, 1–9.

Hendrickson, W. A. (1979). *Phase information from anomalous-scattering measurements. Acta Cryst.* A**35**, 245–247.

Hendrickson, W. A. (1991). *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science*, **254**, 51–58.

Kartha, G. & Parthasarathy, R. (1965*a*). *Combination of multiple isomorphous replacement and anomalous dispersion data for protein structure determination. I. Determination of heavy-atom positions in protein derivatives. Acta Cryst.* **18**, 745–749.

Kartha, G. & Parthasarathy, R. (1965*b*). *Combination of multiple isomorphous replacement and anomalous dispersion data for protein structure determination. II. Correlation of the heavy-atom positions in different isomorphous protein crystals. Acta Cryst.* **18**, 749–753.

Matthews, B. W. (1966*a*). *The determination of the position of the anomalously scattering heavy atom groups in protein crystals. Acta Cryst.* **20**, 230–239.

Matthews, B. W. (1966*b*). *The extension of the isomorphous replacement method to include anomalous scattering measurements. Acta Cryst.* **20**, 82–86.

Matthews, B. W. (1970). *Determination and refinement of phases for proteins*. In *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 146–159. Copenhagen: Munksgaard.

North, A. C. T. (1965). *The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. Acta Cryst.* **18**, 212–216.

Okaya, Y. & Pepinsky, R. (1960). *New developments in the anomalous dispersion method for structure analysis.* In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, pp. 273–299. London: Pergamon Press.

Perutz, M. F. (1956). *Isomorphous replacement and phase determination in non-centrosymmetric space groups. Acta Cryst.* **9**, 867–873.

Ramachandran, G. N. & Raman, S. (1956). *A new method for the structure analysis of non-centrosymmetric crystals. Curr. Sci.* **25**, 348–351.

Ramaseshan, S. (1964). *The use of anomalous scattering in crystal structure analysis*. In *Advanced Methods of Crystallography*, edited by G. N. Ramachandran, pp. 67–95. London: Academic Press.

Rossmann, M. G. (1960). *The accurate determination of the position and shape of heavy-atom replacement groups in proteins. Acta Cryst.* **13**, 221–226.

Rossmann, M. G. (1961). *The position of anomalous scatterers in protein crystals. Acta Cryst.* **14**, 383–388.

Singh, A. K. & Ramaseshan, S. (1966). *The determination of heavy atom positions in protein derivatives. Acta Cryst.* **21**, 279–280.

International Tables for Crystallography (2012). Vol. F, Chapter 14.2, pp. 373–378.

Chapter 14.2. Multiwavelength anomalous diffraction

J. L. Smith and W. A. Hendrickson

Anomalous-scattering effects measured at several X-ray wavelengths can provide a direct solution to the crystallographic phase problem. For many years this was appreciated as a hypothetical possibility (Okaya & Pepinsky, 1956), but, until tunable synchrotron radiation became available, experimental investigation with the weakly diffracting crystals of biological macromolecules was limited to one heroic experiment (Hoppe & Jakubowski, 1975). Multiwavelength anomalous diffraction (MAD) became a dominant phasing method in macromolecular crystallography with the advent of reliable, brilliant synchrotron-radiation sources, the adoption of cryopreservation techniques for crystals of macromolecules, and the development of general anomalous-scatterer labels for proteins and nucleic acids.

Anomalous scattering, first recognized as a source of phase information by Bijvoet (1949), has been employed since the early days of macromolecular crystallography (Blow, 1958). It has been used to locate positions of anomalous scatterers (Rossmann, 1961), to supplement phase information from isomorphous replacement (North, 1965; Matthews, 1966*a*) and to identify the enantiomorph of the heavy-atom partial structure in multiple isomorphous replacement (MIR) phasing (Matthews, 1966*b*). Anomalous scattering at a single wavelength was the sole source of phase information in the structure determination of crambin (Hendrickson & Teeter, 1981), an important precursor to development of MAD. MAD differs from these other applications in using anomalous scattering at several wavelengths for complete phase determination without approximations or simplifying assumptions.

14.2.1. Anomalous scattering factors

The scattering of X-rays by an isolated atom is described by the atomic scattering factor, f^0, based on the assumption that the electrons in the atom oscillate as free electrons in response to X-ray stimulation. The magnitude of f^0 is normalized to the scattering by a single electron. Thus the 'normal' scattering factor f^0 is a real number, equal to the Fourier transform of the electron-density distribution of the atom. At zero scattering angle ($s = \sin\theta/\lambda = 0$), f^0 equals Z, the atomic number. f^0 falls off rapidly with increasing scattering angle due to weak scattering by the diffuse parts of the electron-density distribution. In reality, electrons in an atom do not oscillate freely because they are bound in atomic orbitals. Deviation from the free-electron model of atomic scattering is known as anomalous scattering. Using a classical mechanical model (James, 1948), an atom scatters as a set of damped oscillators with resonant frequencies matched to the absorption frequencies of the electronic shells. The total atomic scattering factor, f, is thus a complex number. f is denoted as a sum of 'normal' and 'anomalous' components, where the anomalous components are corrections to the free-electron model:

$$f = f^0 + f' + if''. \qquad (14.2.1.1)$$

f' and f'' are expressed in electron units, as is f^0. The real component of anomalous scattering, f', is in phase with the normal scattering, f^0, whilst the imaginary component, f'', is out of phase by $\pi/2$.

The imaginary component of anomalous scattering, f'', is proportional to the atomic absorption coefficient of the atom, μ_a, at X-ray energy E:

$$f''(E) = (mc/4\pi e^2\hbar)E\mu_a(E), \qquad (14.2.1.2)$$

where m is the electronic mass, c is the speed of light, e is the electronic charge and h $(= 2\pi\hbar)$ is Planck's constant. Thus, f'' can be determined experimentally by measurement of the atomic absorption coefficient. The relationship between f'' and f' is known as the Kramers–Kronig dispersion relation (James, 1948; Als-Nielsen & McMorrow, 2001):

$$f'(E) = \left(\frac{2}{\pi}\right) P \int\limits_0^\infty \frac{E'f''(E')}{E^2 - E'^2}\,\mathrm{d}E', \qquad (14.2.1.3)$$

where P represents the Cauchy principal value of the integral such that integration over E' is performed from 0 to $(E - \varepsilon)$ and from $(E + \varepsilon)$ to ∞, and then the limit $\varepsilon \to 0$ is taken. The principal value of the integral can be evaluated numerically from limited spectral data that have been scaled to theoretical f'' scattering factors (or μ_a absorption coefficients) at points remote from the absorption edge.

Anomalous scattering is present for all atomic types at all X-ray energies. However, the magnitudes of f' and f'' are negligible at X-ray energies far removed from the resonant frequencies of the atom. This includes all light atoms (H, C, N, O) of biological macromolecules at all X-ray energies commonly used for crystallography. f' and f'' are rather insensitive to scattering angle, unlike f^0, because the electronic resonant frequencies pertain to inner electron shells, which have radii much smaller than the X-ray wavelengths used for anomalous-scattering experiments. The magnitudes of f' and f'' are greatest at X-ray energies very near resonant frequencies, and are also highly energy-dependent (Fig. 14.2.1.1). This property of anomalous scattering is exploited in MAD.

Three means are available for evaluating anomalous scattering factors, f' and f''. Calculations from first principles on isolated elemental atoms are accurate for energies remote from resonant frequencies (Cromer & Liberman, 1970*a*,*b*). However, these calculated values do not apply to the energies most critical in a MAD experiment. f' and f'' can also be estimated by fitting to diffraction data measured at different energies (Templeton *et al.*, 1982). Finally, f'' can be obtained from X-ray absorption spectra by the equation above, and f' from f'' by the Kramers–Kronig transform [equation (14.2.1.3); Hendrickson *et al.*, 1988; Smith, 1998]. Both the precise position of a resonant frequency and the values of f' and f'' near resonance generally depend on transitions to unoccupied molecular orbitals, and are quite sensitive to the electronic environment surrounding the atom. Complexities in the X-ray absorption edge, particularly so-called 'white lines', can enhance the anomalous scattering considerably (Fig. 14.2.1.1). Thus, experimental measurements are needed to select wavelengths for optimal signals, and the values of f' and f'' should be determined either from an absorption spectrum or by refinement against the diffraction data.

X-ray spectra near absorption edges of anomalous scatterers depend on the orientation of the local chemical environment in the X-ray beam, which is polarized for synchrotron radiation. The anisotropy of anomalous scattering may affect both the edge position and the magnitude of absorption. In such cases, f' and f'' for individual atoms are also dependent on orientation. Orientational averaging due to multiple anomalous scatterer sites or crystallographic symmetry may prevent macroscopic detection of polarization effects in crystals. A formalism to describe anisotropic anomalous scattering in which f' and f'' are tensors has been developed by Templeton & Templeton (1988). Fanchon & Hendrickson (1990) have developed a technique to refine f' and f'' tensors against MAD data. Although anomalous scattering labels such as the commonly used selenomethionine and related compounds are strongly anisotropic (Templeton & Templeton, 1988; Hendrickson *et al.*, 1990), anisotropy is generally ignored in MAD for biological macromolecules.

14.2.2. A phase equation for MAD

The impact of anomalous scattering on diffraction measurements can be evaluated by substituting the scattering-factor expression [equation (14.2.1.1)] into the structure-factor equation.

$$ {}^{\lambda}F_{\rm obs}(\mathbf{h}) = \sum_{i=1}^{\text{No. of atoms}} (f^0 + f'_{\lambda} + if''_{\lambda})_i \exp[-B_i s^2(\mathbf{h})] \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_i). \qquad (14.2.2.1) $$

It is convenient to separate terms in the structure-factor expression according to wavelength dependence (Karle, 1980). A wavelength-independent structure factor ${}^{0}F_T$ with phase φ_T is defined to represent the total normal scattering from all atoms (Hendrickson, 1985) as

$$ \begin{aligned} {}^{0}F_T(\mathbf{h}) &= \sum_{i=1}^{\text{No. of atoms}} f_i^0 \exp\left[-B_i s^2(\mathbf{h})\right] \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_i) \\ &= |{}^{0}F_T| \exp(i\varphi_T). \end{aligned} \qquad (14.2.2.2) $$

The anomalous scattering is confined to wavelength-dependent 'anomalous' structure factors ${}^{\lambda}\mathbf{F}'$ and ${}^{\lambda}\mathbf{F}''$, representing the real and imaginary components of anomalous scattering for all atoms. In general, the anomalous structure factors are considered for only the subset of atoms with detectable anomalous scattering, leading to the anomalous structure factors ${}^{\lambda}\mathbf{F}'_A$ and ${}^{\lambda}\mathbf{F}''_A$:

$$ {}^{\lambda}F'_A(\mathbf{h}) = \sum_{j=1}^{N_{\rm ano}} f'_{\lambda j} \exp\left[-B_j s^2(\mathbf{h})\right] \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j) \qquad (14.2.2.3) $$

$$ {}^{\lambda}F''_A(\mathbf{h}) = \sum_{j=1}^{N_{\rm ano}} if''_{\lambda j} \exp\left[-B_j s^2(\mathbf{h})\right] \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j), \qquad (14.2.2.4) $$

where $N_{\rm ano}$ is the number of anomalous scatterers. The anomalous structure factors ${}^{\lambda}F'_A$ and ${}^{\lambda}F''_A$ can also be expressed in terms of the normal structure-factor components, ${}^{0}F_{A_k}$ with phase φ_{A_k}, for the subsets of atoms that comprise each kind of anomalous scatterer (Karle, 1980),

$$ {}^{\lambda}F'_A = \sum_{k=1}^{\text{No. of kinds}} (f'_{\lambda}/f^0)_k \, {}^{0}F_{A_k} \qquad (14.2.2.5) $$

$$ {}^{\lambda}F''_A = \sum_{k=1}^{\text{No. of kinds}} (if''_{\lambda}/f^0)_k \, {}^{0}F_{A_k}. \qquad (14.2.2.6) $$

Thus, the wavelength-dependent experimental structure factor ${}^{\lambda}F_{\rm obs}$ can be represented using normal structure factors only:

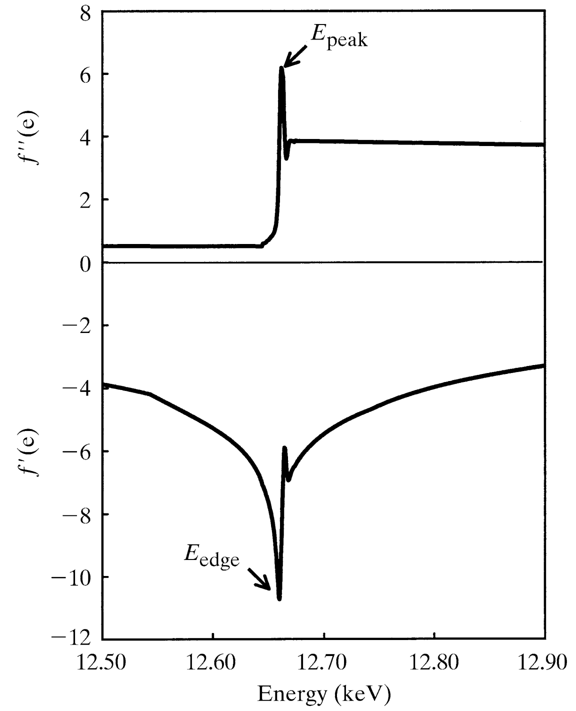


Figure 14.2.1.1
Anomalous scattering factors for Se in a protein labelled with selenomethionine. The spectra are a hybrid of experimental values derived from an absorption spectrum of a SeMet protein for energies near the Se K absorption edge and calculated values for energies remote from the edge. The Se K edge occurs at 12 660 eV, corresponding to a wavelength of 0.9800 Å. Anomalous-scattering effects are enhanced by a white line just above the edge. The position of the absorption edge ($E_{\rm edge}$) is the inflection point of the μ_a (and f'') spectrum, and $E_{\rm peak}$ is the energy of peak absorption just above the edge. These energies correspond to the wavelengths $\lambda_{\rm edge}$ and $\lambda_{\rm peak}$ in a MAD experiment because the magnitudes of f' and f'' are greatest at $E_{\rm edge}$ and $E_{\rm peak}$, respectively.

$$ {}^{\lambda}F_{\rm obs} = {}^{0}F_T + \sum_{k=1}^{\text{No. of kinds}} \left(\frac{f'_{\lambda}}{f^0} + i\frac{f''_{\lambda}}{f^0} \right)_k {}^{0}F_{A_k} \qquad (14.2.2.7) $$

This factorization is convenient because all wavelength dependence is confined to the anomalous scattering factors f' and f'', which are independent of atomic positions, occupancies and thermal parameters. In addition, an electron-density map for the total structure should be based on normal scattering by all atoms, represented by the structure factor ${}^{0}F_T$ with phase φ_T. As described in Section 14.2.1, the normal scattering factor f^0 is strongly dependent on scattering angle, whereas f' and f'' are nearly invariant with s.

A very useful observational equation is obtained by applying the law of cosines to ${}^{0}F_T$, ${}^{\lambda}F'_A$ and ${}^{\lambda}F''_A$ (Karle, 1980; Hendrickson, 1985). Anomalous structure factors are treated separately for each kind of anomalous scatterer. The number of terms in the resulting expression is $(q+1)^2$ for q kinds of anomalous scatterers. For the commonest case of one kind of anomalous scatterer

$$ \begin{aligned} |{}^{\lambda}F_{\rm obs}(\pm\mathbf{h})|^2 &= |{}^{0}F_T|^2 + a_{\lambda}|{}^{0}F_A|^2 \\ &\quad + b_{\lambda}|{}^{0}F_T|\,|{}^{0}F_A|\cos(\varphi_T - \varphi_A) \\ &\quad \pm c_{\lambda}|{}^{0}F_T|\,|{}^{0}F_A|\sin(\varphi_T - \varphi_A), \end{aligned} \qquad (14.2.2.8) $$

where $a_{\lambda} = (f''^2_{\lambda} + f'^2_{\lambda})/(f^0)^2$, $b_{\lambda} = 2f'_{\lambda}/f^0$ and $c_{\lambda} = 2f''_{\lambda}/f^0$. $|{}^{\lambda}F_{\rm obs}(\pm\mathbf{h})|$ refers to the Bijvoet mate reflections $+\mathbf{h}$ and $-\mathbf{h}$. The MAD observational equation illustrates the orthogonal contributions to phasing made by the real (f') and imaginary (f'')

components of anomalous scattering. 'Dispersive' phase information derives from differences between $|F_{\rm obs}|$ at wavelengths having different values of f' and contributes to $\cos(\varphi_T - \varphi_A)$. 'Bijvoet' phase information derives from the Friedel difference $|F_{\rm obs}(+\mathbf{h})| - |F_{\rm obs}(-\mathbf{h})|$ at a wavelength with substantial f'' and contributes to $\sin(\varphi_T - \varphi_A)$. Phase information is enhanced by selection of wavelengths for data measurement that maximize the magnitudes of both f' and f''. Apart from ignoring the very weakest anomalous-scattering effects, the MAD observational equation (14.2.2.8) involves no approximations.

14.2.3. Diffraction ratios for estimating the MAD phasing signal

The first consideration in design of a MAD experiment is the choice of anomalous scatterer(s) with consideration of the magnitude of the phasing signal. Estimation of the total scattering by the macromolecule and the potential phasing signal generated by the anomalous scatterer(s) under consideration is informative.

The magnitude of the MAD phasing signal is estimated as the ratio of the expected dispersive or Bijvoet difference to the expected total scattering of the macromolecule. This is based on calculation of the expected root-mean-square structure amplitude (rms$|F|$) (Wilson, 1942),

$$\mathrm{rms}|F| = \langle |F|^2 \rangle^{1/2} = \left(\textstyle\sum f_i^2\right)^{1/2} = N^{1/2} f \qquad (14.2.3.1)$$

for N identical atoms. The expected total scattering of the macromolecule is estimated at $s = 0$ using an average non-hydrogen atom. Based on atomic frequencies in biological macromolecules, the average values of f^0 are 6.70 e for proteins, 7.20 e for DNA and 7.26 e for RNA. The average number of non-hydrogen atoms and molecular mass per residue are 7.7 atoms and 110 Da for proteins, 21.8 atoms and 292 Da for DNA, and 22.4 atoms and 304 Da for RNA. These averages result in the following expressions for estimated total scattering of biological macromolecules:

$$\begin{aligned}
\mathrm{rms}|^0F_T|_{\rm protein} &\approx 6.70(\text{No. of atoms})^{1/2} \approx (346 \times \text{No. of amino acids})^{1/2}\\
&\approx (3.14 \times \text{molecular mass})^{1/2}\\
\mathrm{rms}|^0F_T|_{\rm DNA} &\approx 7.20(\text{No. of atoms})^{1/2} \approx (1128 \times \text{No. of nucleotides})^{1/2}\\
&\approx (3.87 \times \text{molecular mass})^{1/2}\\
\mathrm{rms}|^0F_T|_{\rm RNA} &\approx 7.26(\text{No. of atoms})^{1/2} \approx (1183 \times \text{No. of nucleotides})^{1/2}\\
&\approx (3.89 \times \text{molecular mass})^{1/2}. \qquad (14.2.3.2)
\end{aligned}$$

Note: the estimated total scattering of a protein is coincidentally $\approx (\pi \times \text{molecular mass})^{1/2}$.

The diffraction ratios relevant to a MAD experiment with N anomalous-scatterer sites are

$$\frac{\mathrm{rms}\big||^{\lambda1}F_{\rm obs}| - |^{\lambda2}F_{\rm obs}|\big|}{\mathrm{rms}|^0F_T|} \approx (N/2)^{1/2}\,\frac{|f'_{\lambda1} - f'_{\lambda2}|}{\mathrm{rms}|^0F_T|} \qquad (14.2.3.3)$$

for the dispersive signal and

$$\frac{\mathrm{rms}\big||^{\lambda}F^+_{\rm obs}| - |^{\lambda}F^-_{\rm obs}|\big|}{\mathrm{rms}|^0F_T|} \approx (N/2)^{1/2}\,\frac{2f''_{\lambda}}{\mathrm{rms}|^0F_T|} \qquad (14.2.3.4)$$

for the Bijvoet signal. The diffraction ratios, analogous to similar relations for isomorphous replacement (Crick & Magdoff, 1956), are equivalent to the expected fractional changes in intensity due to anomalous scattering, and, as such, can be compared directly to the $R_{\rm sym}$ estimate of error in the experimental data for evaluation of the phasing signal. Of course, the phasing signal may be diminished by partial occupancy or thermal motion, as for normal scattering.

14.2.4. Experimental considerations

The design and execution of a MAD experiment are distinguished from monochromatic experiments in macromolecular crystallography primarily by the stringent criteria for wavelength selection.

The largest MAD phasing signal is obtained at energies with the most extreme values of f' and f'', which correspond to the sharpest features of the absorption edge (Fig. 14.2.1.1). The energy of peak absorption just above the edge ($E_{\rm peak}$) corresponds to the wavelength of maximum f'' and optimal Bijvoet signal ($\lambda_{\rm peak}$). Typically, the orthogonal dispersive signal is optimized by recording one data set at the wavelength corresponding to the inflection point of the absorption edge (minimum f', $\lambda_{\rm edge}$), and one or more data sets at remote wavelengths having f' with smaller magnitudes ($\lambda_{\rm remote}$). The choice of the remote wavelength(s) is experiment dependent. If only one remote wavelength is used, it is typically on the high-energy side of the absorption edge due to the larger Bijvoet signal. The remote wavelength(s) may also be chosen to avoid complications from other edges or to obtain data at a wavelength optimal for model refinement. In the case of anomalous scatterers that exhibit sharp 'white line' features, the dispersive signal may be optimized between the minimum of f' at the ascending edge ($\lambda_{\rm edge}$) and the local maximum of f' at the descending inflection point ($\lambda_{\rm descent}$).

The features of an X-ray absorption edge are in many cases very sharp, with the energies of the inflection point and peak absorption separated by as little as 2 eV. Therefore, it is critical to determine $\lambda_{\rm edge}$ and $\lambda_{\rm peak}$ experimentally by recording the absorption edge from the labelled macromolecule at the time of a MAD experiment. Even when the position of the edge is well known, small unanticipated chemical changes in the sample or calibration errors in the X-ray beam can reduce the MAD signal very significantly.

The MAD phasing signal is derived from intensity differences that may be similar in magnitude to measurement errors. Thus a general philosophy in the design of a MAD experiment is to equalize systematic errors among the measurements whose differences will contribute to each phase determination. This is achieved for each unique reflection by recording Bijvoet measurements at all wavelengths from the same asymmetric portion of diffraction space at nearly the same time. If crystal decay necessitates use of multiple crystals in a MAD experiment, blocks of Bijvoet data should be recorded identically at all of the selected wavelengths from each crystal contributing to the data set. Bijvoet mates can be recorded simultaneously by alignment of the crystal with a mirror plane of diffraction symmetry perpendicular to the rotation axis, or Friedel images can be recorded in an 'inverse beam' experiment. Inverse-beam geometry is a hypothetical method for measurement of Friedel data using both the forward and reverse directions of the incident X-ray beam. In a real experiment, diffraction images and their Friedel equivalents are recorded at crystal positions related by 180° rotation about any axis perpendicular to the incident beam, usually the data-collection axis. The inverse-beam experiment requires neither crystal symmetry nor crystal alignment, and is well suited to crystals mounted in random orientations.

The multiwavelength measurements for each unique reflection will be identically redundant and have nearly equal systematic errors if identical blocks of Bijvoet data are collected, as described above. When such a data-collection strategy is followed, the resulting MAD data set will include all multiwavelength Bijvoet measurements for all regions of the reciprocal lattice that are covered in the experiment.

Cryopreservation of crystals is of enormous benefit to MAD. Systematic error due to radiation damage is eliminated or greatly diminished. Systematic differences between crystals are eliminated in cases where a complete MAD data set is measured from a single frozen crystal. Intensities of weak reflections are estimated more accurately because less material contributes to diffuse background scatter in the mounts used for frozen crystals than for unfrozen crystals.

Measurement errors are of major importance in all areas of macromolecular crystallography, but are the limiting factor in phase determination by MAD. MAD data should be of high quality by the usual measures ($R_{\rm sym}$, redundancy, completeness), especially in experiments where the phasing signal is weak. Good counting statistics are of paramount importance. Experimental error, estimated by $R_{\rm sym}$, increases with increasing scattering angle because of the strong fall-off of f^0 with s. In a carefully designed experiment, the effect of increasing $R_{\rm sym}$ with s is mitigated somewhat by equalizing systematic errors and by averaging highly redundant data. Disappearance of the phasing signal into $R_{\rm sym}$ noise is the major reason that useful MAD phases are not obtained to the diffraction limit of crystals, even though anomalous scattering does not diminish with increasing s.

The optimal number of data-collection wavelengths for successful phase determination by MAD has been debated. In most cases, it is necessary to measure data at $\lambda_{\rm edge}$, $\lambda_{\rm peak}$ and a $\lambda_{\rm remote}$ in order to take advantage of the most extreme values of f' and f''. If f' values at $\lambda_{\rm edge}$ and $\lambda_{\rm peak}$ are different enough to produce a detectable dispersive signal, then phases can be obtained from three measurements: $|F^+|$ and $|F^-|$ at $\lambda_{\rm peak}$, and either $|F^+|$ or $|F^-|$ at $\lambda_{\rm edge}$. However, redundancy is one of the best ways to minimize the effects of measurement error in macromolecular crystallography. Redundant Bijvoet signals can be obtained at $\lambda_{\rm peak}$ and at any $\lambda_{\rm remote}$ above the absorption edge if both $|F^+|$ and $|F^-|$ are measured at each wavelength. Likewise, the dispersive signal between measurements at $\lambda_{\rm edge}$ and $\lambda_{\rm remote}$ is also redundant if both $|F^+|$ and $|F^-|$ measurements are taken at each wavelength. More highly redundant four- or five-wavelength MAD experiments may be advantageous, although greater redundancy should not be gained at the cost of good counting statistics. Brilliant synchrotron sources and high-speed detectors make rapid measurement of complete multiwavelength data sets possible, but the practical feasibility is often compromised by radiation damage. Phase information from the Bijvoet signal at a single wavelength (preferably $\lambda_{\rm peak}$) can also be used as a basis for structure determination. The phase probability distribution from single-wavelength anomalous scattering is in general bimodal, and must be resolved with additional phase information. This could be the partial structure of anomalous scatterers, as in the classic crambin experiment (Hendrickson & Teeter, 1981), or the real-space constraints, such as solvent flattening or redundancy averaging, that are applied in common schemes for phase refinement by density modification.

14.2.5. Data handling

Two general approaches to data handling for MAD have been employed.

An extreme interpretation of the scheme for equalizing systematic errors is known as 'phase first, merge later' (Hendrickson, 1985; Hendrickson & Ogata, 1997). The idea is that systematic errors may be amplified by merging data, and that this may obscure a weak phasing signal. In this approach, the individual observations constituting a multiwavelength Bijvoet set, as determined by the data-collection strategy, are grouped together and scaled. There may be redundant multiwavelength sets of observations, but these are merged only after individual phase evaluations have been made. Error estimates from the phasing, or the agreement of redundant phase determinations, can be incorporated into weights for averaging, or can be used to reject outliers. Complicated, experiment-dependent bookkeeping is required to assemble exactly the correct observations into each unmerged set of multiwavelength measurements. However, the 'phase first, merge later' approach may be advantageous for MAD data sets from multiple crystals, or when minor disasters disrupt the experiment and thwart the data-collection strategy.

A second approach, known as 'merge first, phase later', is to scale and merge data at each wavelength, keeping Bijvoet pairs separate, and then to scale data at all wavelengths to one another (Ramakrishnan & Biou, 1997). The idea is that the multiwavelength Bijvoet measurements are identically redundant for each unique reflection if the MAD data were measured according to the strategy outlined in Section 14.2.4. Thus, merging the redundancies should reduce systematic errors in the amplitude differences used for phasing. The 'merge first, phase later' approach is computationally simpler than the 'phase first, merge later' approach because it is experiment independent. However, unanticipated experimental disasters may be more difficult to overcome in the 'merge first, phase later' approach to data handling.

Of course, if the MAD signal is strong relative to the experimental error, either approach to data handling should be successful. Data scaling in both approaches may be done most easily and reliably by scaling all data against a standard data set, such as the unique data from one wavelength with Bijvoet mates averaged. In general a dogmatic approach to data handling is best avoided in favour of whichever computational technique or combination of techniques is most suited to the problem at hand. Factors such as the strength of the MAD signal, data-collection strategy, number of crystals contributing to the data set, crystal quality and experimental disasters should be taken into account.

14.2.6. Approaches to MAD phasing

There are two general approaches to MAD phasing. In the explicit approach, the MAD observational equation is solved directly (Hendrickson *et al.*, 1988; Hendrickson & Ogata, 1997). In the pseudo-MIR approach, MAD phasing is treated as a special case of multiple isomorphous replacement (Burling *et al.*, 1996; Terwilliger, 1997; Ramakrishnan & Biou, 1997). Both approaches have been quite successful, and each has advantages and disadvantages. For complete phase determination by either method, the partial structure of the anomalous scatterers must be determined. The explicit and pseudo-MIR approaches differ in when the partial structure is determined and in how it is refined.

The explicit approach provides the quantities $|{}^0F_T|$, $|{}^0F_A|$ and $(\varphi_T - \varphi_A)$ by direct fit of the $|{}^\lambda F_{\rm obs}|$ to the MAD observational equation (14.2.2.8). No anomalous-scatterer partial structure model is required in this first step of phasing. Estimates of the anomalous scattering factors at the wavelengths of data collection are required. These estimates can be refined (Weis *et al.*, 1991), so they need not be highly accurate. Redundancies are merged to produce a unique data set at the level of the derived quantities $|^0F_T|$, $|^0F_A|$, $(\varphi_T - \varphi_A)$ and their error estimates. The anomalous-scatterer partial structure is determined from the derived estimates of $|^0F_A|$ and refined against these amplitudes. In the second step of phasing, φ_T is derived from the phase difference $(\varphi_T - \varphi_A)$ and weights are calculated for a Fourier synthesis from $|^0F_T|$ and φ_T. Phase probability distributions (*ABCD* coefficients; Hendrickson & Lattman, 1970) derived from the MAD observational equation (14.2.2.8) can be used directly in the explicit approach (Pähler *et al.*, 1990). A probabilistic treatment based on maximum likelihood theory has also been developed (de La Fortelle & Bricogne, 1997). There are two advantages to the explicit approach. First, it is amenable to the 'phase first, merge later' scheme of data handling because refinement of the anomalous-scatterer partial structure is entirely separate from phase calculation. The second principal advantage of the explicit approach is the calculation of an experimentally derived estimate of the normal structure amplitude $|^0F_A|$ for the anomalous scatterer. This is the quantity with which the partial structure of anomalous scatterers is most directly solved and refined. However, extraction of reliable $|^0F_A|$ estimates from data with low signal-to-noise can be difficult. Bayesian methods of $|^0F_A|$ estimation (Terwilliger, 1994; Krahn *et al.*, 1999) have been shown to be more robust than least-squares methods.

In the pseudo-MIR approach, data at one wavelength are designated as 'native' data, which include anomalous scattering, and data at the other wavelengths as 'derivative' data. This approach has the advantage that nothing need be known about the anomalous scattering factors at any time during phasing. These quantities are incorporated into heavy-atom atomic 'occupancies' and refined along with other parameters. Of course, the partial structure of anomalous scatterers must be known, and its refinement is concurrent with phasing. This may be a principal advantage of the pseudo-MIR approach, because the anomalous-scatterer parameter refinement may be more reliable when incorporated into phasing than when done against $|^0F_A|$ estimates. Greater weight is given to the data set selected as 'native' in refinement of the 'heavy-atom' parameters in some implementations of the pseudo-MIR approach, although others treat data at all wavelengths equivalently (Terwilliger & Berendzen, 1997). The amplitudes $|^0F_A|$ are not a by-product of the pseudo-MIR approach.

14.2.7. Determination of the anomalous-scatterer partial structure

Determination of the partial structure of anomalous scatterers is a prerequisite for MAD-phased electron density, regardless of the phasing technique. As described above, the optimal quantities for solving and refining the partial structure of anomalous scatterers are the normal structure amplitudes $|^0F_A|$. Frequently $|^0F_A|$ values are not extracted from the MAD measurements, and the largest Bijvoet or dispersive differences are used instead. This involves the approximation of representing structure amplitudes $(|^0F_A|)$ as the subset of larger differences $(||F^+| - |F^-||$ or $||F_{\lambda^1}| - |F_{\lambda^2}||)$. The approximation is accurate for only a small fraction of reflections because there is little correlation between φ_A and φ_T. However, it suffices for a suitably strong signal and a suitably small number of sites. Patterson methods are quite successful in locating anomalous scatterers when the number of sites is small. However, the aim of MAD is to solve the macromolecule structure from one MAD data set using any number of anomalous scatterer sites. For larger numbers of sites, statistical direct methods may be employed.

The correct enantiomorph for the anomalous-scatterer partial structure also must be determined (φ_A *versus* $-\varphi_A$) in order to obtain an electron-density image of the macromolecule. However, it cannot be determined directly from MAD data. The correct enantiomorph is chosen by comparison of electron-density maps based on both enantiomorphs of the partial structure. Unlike the situation for pure MIR, the density based on the incorrect enantiomorph of the anomalous-scatterer partial structure is not the mirror image of that based on the correct enantiomorph and contains no image of the macromolecule. The correct map is distinguished by features such as a clear solvent boundary, positive correlation of redundant densities and a macromolecule-like density histogram. If the anomalous-scattering centres form a centric array, then the two enantiomorphs are identical and both maps are correct.

14.2.8. General anomalous-scatterer labels for biological macromolecules

MAD requires a suitable anomalous scatterer, of which none are generally present in naturally occurring proteins or nucleic acids. However, selenomethionine (SeMet) substituted for the natural amino acid methionine (Met) is a general anomalous-scattering label for proteins (Hendrickson, 1985), and is the anomalous scatterer most frequently used in MAD. The *K* edge of Se is the most accessible for MAD experiments ($\lambda = 0.98$ Å).

The SeMet label is especially general and convenient because it is introduced by biological substitution of SeMet for methionine. This is achieved by blocking methionine biosynthesis and substituting SeMet for Met in the growth medium of the cells in which the protein is produced. Production of SeMet protein in bacteria is generally straightforward (Hendrickson *et al.*, 1990; Doublié, 1997) and has also been accomplished in eukaryotic cells (Lustbader *et al.*, 1995; Bellizzi *et al.*, 1999).

Methionine is a particularly attractive target for anomalous-scatterer labelling because the side chain is usually buried in the hydrophobic core of globular proteins where it is relatively better ordered than are surface side chains. The labelling experiment provides direct evidence for isostructuralism of Met and SeMet proteins. All proteins in the biological expression system have SeMet substituted for Met at levels approaching 100%. The cells are viable, therefore the proteins are functional and isostructural with their unlabelled counterparts to the extent required by function.

The natural abundance of methionine in soluble proteins is approximately one in fifty amino acids, providing a typical MAD phasing signal of 4–6% of $|F|$ [equations (14.2.3.3) and (14.2.3.4)]. Typical extreme values for the anomalous scattering factors are $f'_{\rm min} \sim -10$ e and $f''_{\rm max} \sim 6$ e (Fig. 14.2.1.1). SeMet is more sensitive to oxidation than is Met, and care must be taken to maintain a homogeneous oxidation state. Generally, the reduced state is maintained by addition of disulfide reducing agents to SeMet protein and crystals. However, the oxidized forms of Se have sharper *K*-edge features and f' and f'' values of greater magnitude than does the reduced form (Smith & Thompson, 1998). This property has been exploited to enhance anomalous signals by intentional oxidation of SeMet protein (Sharff *et al.*, 2000). SeMet is also a useful isomorphous-replacement label with a signal of ~10% of $|F|$. Prior knowledge of the sites of labelling is extremely useful during initial fitting of a protein sequence to electron density. Also, noncrystallographic symmetry operators can usually be defined more reliably from Se positions in SeMet protein than by heavy-atom positions in MIR due to the uniformity and completeness of labelling (Tesmer *et al.*, 1996).

An analogous general label is available for nucleic acids in the form of brominated bases, particularly 5-bromouridine, which is isostructural with thymidine. The *K* edge of Br corresponds to a wavelength of 0.92 Å, which is quite favourable for data collection.

References

Als-Nielsen, J. & McMorrow, D. F. (2001). *Elements of Modern X-ray Physics*. New York: John Wiley & Sons.

Bellizzi, J. J. III, Widom, J., Kemp, C. W. & Clardy, J. (1999). *Producing selenomethionine-labeled proteins with a baculovirus expression vector system. Structure*, **7**, R263–R267.

Bijvoet, J. M. (1949). *Phase determination in direct Fourier-synthesis of crystal structures. Proc. Acad. Sci. Amst.* B**52**, 313–314.

Blow, D. M. (1958). *The structure of haemoglobin. VII. Determination of phase angles in the non-centrosymmetric [100] zone. Proc. R. Soc. London Ser. A*, **247**, 302–335.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. Science*, **271**, 72–77.

Crick, F. H. C. & Magdoff, B. S. (1956). *The theory of the method of isomorphous replacement for protein crystals. I. Acta Cryst.* **9**, 901–908.

Cromer, D. T. & Liberman, D. (1970*a*). *Relativistic calculation of anomalous scattering factors for X-rays. J. Chem. Phys.* **53**, 1891–1898.

Cromer, D. T. & Liberman, D. (1970*b*). *Relativistic calculation of anomalous scattering factors for X-rays.* Report LA-4403. Los Alamos National Laboratory, USA.

Doublié, S. (1997). *Preparation of selenomethionyl proteins for phase determination*. *Methods Enzymol.* **276**, 523–530.

Fanchon, E. & Hendrickson, W. A. (1990). *Effect of the anisotropy of anomalous scattering on the MAD phasing method. Acta Cryst.* A**46**, 809–820.

Hendrickson, W. A. (1985). *Analysis of protein structure from diffraction measurement at multiple wavelengths. Trans. Am. Crystallogr. Assoc.* **21**, 11–21.

Hendrickson, W. A., Horton, J. R. & LeMaster, D. M. (1990). *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. EMBO J.* **9**, 1665–1672.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Hendrickson, W. A. & Ogata, C. M. (1997). *Phase determination from multiwavelength anomalous diffraction measurements. Methods Enzymol.* **276**, 494–523.

Hendrickson, W. A., Smith, J. L., Phizackerley, R. P. & Merritt, E. A. (1988). *Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. Proteins Struct. Funct. Genet.* **4**, 77–88.

Hendrickson, W. A. & Teeter, M. M. (1981). *Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. Nature (London)*, **290**, 107–113.

Hoppe, W. & Jakubowski, U. (1975). *The determination of phases of erythrocruorin using the two-wavelength method with iron as anomalous scatterer.* In *Anomalous Scattering*, edited by S. Ramaseshan & S. C. Abrahams, 3–11. Copenhagen: Munksgaard.

James, R. W. (1948). *The Optical Principles of the Diffraction of X-rays.* Reprinted (1982) Ox Bow Press, Woodbridge, CT.

Karle, J. (1980). *Some developments in anomalous dispersion for the structural investigation of macromolecular systems in biology. Int. J. Quantum Chem. Quantum Biol. Symp.* **7**, 357–367.

Krahn, J. M., Sinha, S. & Smith, J. L. (1999). *Successes and prospects for SeMet MAD and large structures. Trans. Am. Crystallogr. Assoc.* **35**, 27–38.

La Fortelle, E. de & Bricogne, G. (1997). *Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol.* **276**, 472–494.

Lustbader, J. W., Wu, H., Birken, S., Pollak, S., Kolks-Gawinowicz, M. A., Pound, A. M., Austen, D., Hendrickson, W. A. & Canfield, R. E. (1995). *The expression, characterization and crystallization of wild-type and selenomethionyl human chorionic gonadotropin. Endocrinology*, **136**, 640–650.

Matthews, B. W. (1966*a*). *The extension of the isomorphous replacement method to include anomalous scattering measurements. Acta Cryst.* **20**, 82–86.

Matthews, B. W. (1966*b*). *The determination of the position of anomalously scattering heavy atom groups in protein crystals. Acta Cryst.* **20**, 230–239.

North, A. C. T. (1965). *The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. Acta Cryst.* **18**, 212–216.

Okaya, Y. & Pepinsky, R. (1956). *New formulation and solution of the phase problem in X-ray analysis of noncentric crystals containing anomalous scatterers. Phys. Rev.* **103**, 1645–1647.

Pähler, A., Smith, J. L. & Hendrickson, W. A. (1990). *A probability representation for phase information from multiwavelength anomalous dispersion. Acta Cryst.* A**46**, 537–540.

Ramakrishnan, V. & Biou, V. (1997). *Treatment of multiwavelength anomalous diffraction data as a special case of multiple isomorphous replacement. Methods Enzymol.* **276**, 538–557.

Rossmann, M. G. (1961). *The position of anomalous scatterers in protein crystals. Acta Cryst.* **14**, 383–388.

Sharff, A. J., Koronakis, E., Luisi, B. & Koronakis, V. (2000). *Oxidation of selenomethionine: some MADness in the method! Acta Cryst.* D**56**, 785–788.

Smith, J. L. (1998). *Multiwavelength anomalous diffraction in macromolecular crystallography.* In *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 221–225. The Netherlands: CCLRC.

Smith, J. L. & Thompson, A. (1998). *Reactivity of selenomethionine – dents in the magic bullet? Structure*, **15**, 815–819.

Templeton, L. K. & Templeton, D. H. (1988). *Biaxial tensors for anomalous scattering of X-rays in selenolanthionine. Acta Cryst.* A**44**, 1045–1051.

Templeton, L. K., Templeton, D. H., Phizackerley, R. P. & Hodgson, K. O. (1982). *L_3-edge anomalous scattering by gadolinium and samarium measured at high resolution with synchrotron radiation. Acta Cryst.* A**38**, 74–78.

Terwilliger, T. C. (1994). *MAD phasing: Bayesian estimates of F_A. Acta Cryst.* D**50**, 11–16.

Terwilliger, T. C. (1997). *Multiwavelength anomalous diffraction phasing of macromolecular structures: analysis of MAD data as single isomorphous replacement with anomalous scattering data using the MADMRG program. Methods Enzymol.* **276**, 530–537.

Terwilliger, T. C. & Berendzen, J. (1997). *Bayesian correlated MAD phasing. Acta Cryst.* D**53**, 571–579.

Tesmer, J. J. G., Klem, T. J., Deras, M. L., Davisson, V. J. & Smith, J. L. (1996). *The crystal structure of GMP synthetase reveals a novel catalytic triad and is a structural paradigm for two enzyme families. Nat. Struct. Biol.* **3**, 74–86.

Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W. A. (1991). *Structure of the calcium-dependent lectin domain from a rat mannose-binding protein determined by MAD phasing. Science*, **254**, 1608–1615.

Wilson, A. J. C. (1942). *Determination of absolute from relative X-ray intensity data. Nature (London)*, **150**, 151–152.

International Tables for Crystallography (2012). Vol. F, Chapter 14.3, pp. 379–383.

Chapter 14.3. Automated MAD and MIR structure solution

T. C. Terwilliger and J. Berendzen

14.3.1. Introduction

In favourable cases, structure solution by X-ray crystallography using the multiwavelength anomalous diffraction (MAD) or multiple isomorphous replacement (MIR) methods can be a straightforward, though often lengthy, process. The recently developed *Solve* software (Terwilliger & Berendzen, 1999*b*) is designed to fully automate this class of structure solution. The overall approach is to link together all the analysis steps that a crystallographer would normally carry out into a seamless procedure, and in the process to convert each decision-making step into an optimization problem.

In the case of both MAD and MIR data, a key element of the procedure is the scoring and ranking of possible solutions. This scoring procedure makes it possible to treat structure solution as an optimization procedure, rather than a decision-making one. In the case of MAD data, a second key element of the procedure is the conversion of MAD data to a pseudo-SIRAS form (Terwilliger, 1994*b*) that allows much more rapid analysis than one involving the full MAD data set.

14.3.2. MAD and MIR structure solution

The MAD and MIR approaches to structure solution are conceptually very similar and share several important steps. Two of these are the identification of possible locations of heavy or anomalously scattering atoms and an analysis of the quality of each of these potential heavy-atom solutions. In each method, trial partial structures for these heavy or anomalously scattering atoms are often obtained by inspection of difference Patterson functions or by semi-automated analysis (*e.g.* Terwilliger *et al.*, 1987; Chang & Lewis, 1994; Vagin & Teplyakov, 1998). In other cases, direct-methods approaches have been used to find heavy-atom sites (Sheldrick, 1990; Miller *et al.*, 1994). Potential heavy-atom solutions found in any of these approaches are often just a starting point for structure solution, with additional sites found by difference Fourier or other approaches.

The analysis of the quality of potential heavy-atom solutions is also very similar in the MIR and MAD methods. In both cases a partial structure is used to calculate native phases for the entire structure, and the electron density that results is examined to see if the expected features of the macromolecule are found. Additionally, the agreement of the heavy-atom model with the difference Patterson function and the figure of merit of phasing are commonly used to evaluate the quality of a solution. In many cases, an analysis of heavy-atom sites by sequential deletion of individual sites or derivatives is often an important criterion of quality as well (Dickerson *et al.*, 1961).

14.3.3. Decision making and structure solution

The process of structure solution can be thought of largely as a decision-making process. In the early stages of solution, a crystallographer must choose which of several potential trial solutions may be worth pursuing. At a later stage, the crystallographer must choose which peaks in a heavy-atom difference Fourier are to be included in the heavy-atom model, and which hand of the solution is correct. At a final stage, the crystallographer must decide whether the solution process is complete and which of the possible heavy-atom models is the best. The most important feature of the *Solve* software is the use of a consistent scoring algorithm as the basis for making all these decisions.

14.3.4. The need for rapid refinement and phasing during automated structure solution

In order to make automated structure solution practical, it was necessary to be able to evaluate heavy-atom solutions very rapidly. This is because the automated approach used by *Solve* requires analysis of many heavy-atom solutions (typically 300–1000). For each heavy-atom solution examined, the heavy-atom sites have to be refined and phases calculated. In implementing automated structure solution, it was important to recognize the need for a trade-off between the most accurate heavy-atom refinement and phasing at all stages of structure solution and the time required to carry it out. The balance chosen for *Solve* was to use the most accurate available methods for final phase calculations, and to use approximate but much faster methods for all refinements and phase calculations. The refinement method chosen on this basis was origin-removed Patterson refinement (Terwilliger & Eisenberg, 1983), which treats each derivative in an MIR data set independently and which is very fast because it does not require phase calculation. The phasing approach used for MIR data thoughout *Solve* is Bayesian correlated phasing (Terwilliger & Berendzen, 1996; Terwilliger & Eisenberg, 1987), which takes into account the correlation of non-isomorphism among derivatives without substantially slowing down phase calculations.

For MAD data, Bayesian calculations of phase probabilities are very slow (*e.g.* Terwilliger & Berendzen, 1997; de La Fortelle & Bricogne, 1997). Consequently, we have used an alternative procedure for all MAD phase calculations except those done at the very final stage. This alternative is to convert the MAD data set into a form that is similar to one obtained in the single isomorphous replacement with anomalous scattering (SIRAS) method. In this way, a single data set with isomorphous and anomalous differences is obtained that can be used in heavy-atom refinement by the origin-removed Patterson refinement method and in phasing by conventional SIRAS phasing (Terwilliger & Eisenberg, 1987).

14.3.5. Conversion of MAD data to a pseudo-SIRAS form

The conversion of MAD data to a pseudo-SIRAS form that has almost the same information content requires two important assumptions. The first assumption is that the structure factor corresponding to anomalously scattering atoms in a structure varies in magnitude but not in phase at various X-ray wavelengths. This assumption will hold when there is one dominant type of anomalously scattering atom. The second is that the structure factor corresponding to anomalously scattering atoms is small compared to the structure factor from all other atoms. As long as these two assumptions hold, the information in a MAD experiment is largely contained in just three quantities: a structure factor (F_o) corresponding to the scattering from non-anomalously scattering atoms, a dispersive or isomorphous difference at a standard wavelength λ_o ($\Delta^{\rm ISO}_{\lambda_o}$), and an anomalous difference ($\Delta^{\rm ANO}_{\lambda_o}$) at the same standard wavelength (Terwilliger, 1994*b*). It is easy to see that these three quantities could be treated just like a SIRAS data set with the 'native' structure factor F_P replaced by F_o, the derivative structure factor F_{PH} replaced by $F_o + \Delta^{\rm ISO}_{\lambda_o}$, and the anomalous difference replaced by $\Delta^{\rm ANO}_{\lambda_o}$ (Terwilliger, 1994*b*). This is the approach taken by *Solve*. In this section, it is briefly shown how these three quantities can be estimated from MAD data.

For a particular reflection and a particular wavelength λ_j, we can write the total normal (*i.e.*, non-anomalous) scattering from a structure ($\mathbf{F}_{{\rm tot},\,\lambda_j}$) as the sum of two components. One is the scattering from all non-anomalously scattering atoms ($\mathbf{F}_o$). This scattering is wavelength-independent. The second is the normal scattering from anomalously scattering atoms ($\mathbf{F}_{H_{\lambda_j}}$) at wavelength λ_j. This term includes wavelength-dependent dispersive shifts in atomic scattering due to the f' term in the scattering factor, but not the anomalous part due to the f'' term. The magnitude of the total scattering factor can then be written in the form

$$F_{{\rm tot},\,\lambda_j} = |\mathbf{F}_o + \mathbf{F}_{H_{\lambda_j}}|. \tag{14.3.5.1}$$

Here $\mathbf{F}_o$ and $\mathbf{F}_{{\rm tot},\,\lambda_j}$ can be thought of corresponding, respectively, to the native structure factor, F_P, and the derivative structure factor, F_{PH}, as used in the method of isomorphous replacement (Blundell & Johnson, 1976). If the scattering from anomalously scattering atoms is small compared to that from all other atoms, equation (14.3.5.1) can be rewritten in the approximate form

$$F_{{\rm tot},\,\lambda_j} \simeq F_o + F_{H_{\lambda_j}} \cos(\alpha), \tag{14.3.5.2}$$

where α is the phase difference between the structure factors corresponding to non-anomalously and anomalously scattering atoms in the unit cell, $\mathbf{F}_o$ and $\mathbf{F}_{H_{\lambda_j}}$, respectively, at this X-ray wavelength.

The data in a MAD experiment consist of observations of structure-factor amplitudes for Bijvoet pairs, $F^+_{\lambda_j}$ and $F^-_{\lambda_j}$, for several X-ray wavelengths λ_j. These can be rewritten in terms of an average structure-factor amplitude $\overline{F}_{\lambda_j}$ and an anomalous difference $\Delta^{\rm ANO}_{\lambda_j}$ (*cf.* Blundell & Johnson, 1976). We would like to convert these into estimates of the amplitude of the structure factor corresponding to the non-anomalously scattering atoms alone, the amplitude of the structure factor corresponding to the entire structure at a standard wavelength, and the anomalous difference at the standard wavelength.

The normal scattering due to anomalously scattering atoms ($\mathbf{F}_{H_{\lambda_j}}$) changes in magnitude but not direction as a function of X-ray wavelength. We can therefore write (Terwilliger, 1994*b*)

$$\mathbf{F}_{H_{\lambda_j}} = \mathbf{F}_{H_{\lambda_o}} \frac{f_o + f'(\lambda_j)}{f_o + f'(\lambda_o)}, \tag{14.3.5.3}$$

where λ_o is an X-ray wavelength arbitrarily defined as a standard, and the real part of the scattering factor for the anomalously scattering atoms at wavelength λ_o is $f_o + f'(\lambda_j)$. A corresponding approximation for the anomalous differences at various wavelengths can also be written (Terwilliger & Eisenberg, 1987)

$$\Delta^{\rm ANO}_{\lambda_j} = \Delta^{\rm ANO}_{\lambda_o} \frac{f''(\lambda_j)}{f''(\lambda_o)}, \tag{14.3.5.4}$$

where $f''(\lambda_j)$ is the imaginary part of the scattering factor for the anomalously scattering atoms at wavelength λ_j. Based on equation (14.3.5.4), anomalous differences at any wavelength can be estimated using measurements at the standard wavelength.

An estimate of the structure-factor amplitude (F_o) corresponding to the scattering from non-anomalously scattering atoms and of the dispersive difference at standard wavelength λ_o ($\Delta^{\rm ISO}_{\lambda_o}$) can be obtained from average structure-factor amplitudes ($\overline{F}_{\lambda_j}$) at any pair of wavelengths λ_i and λ_j by proceeding in two steps. Using equations (14.3.5.2) and (14.3.5.3), the component of $\mathbf{F}_{H_{\lambda_o}}$ along $\mathbf{F}_o$, which we term $\Delta^{\rm ISO}_{\lambda_o}$, can be estimated as

$$\Delta^{\rm ISO}_{\lambda_o} \simeq F_{H_{\lambda_o}} \cos(\alpha) \tag{14.3.5.5}$$

or

$$\Delta^{\rm ISO}_{\lambda_o} \simeq (\overline{F}_{\lambda_i} - \overline{F}_{\lambda_j}) \frac{f_o + f'(\lambda_o)}{f'(\lambda_i) - f'(\lambda_j)}. \tag{14.3.5.6}$$

Then, in turn, this estimate of $\Delta^{\rm ISO}_{\lambda_o}$ can be used to obtain F_o:

$$F_o \simeq \overline{F}_{\lambda_j} - \Delta^{\rm ISO}_{\lambda_o} \frac{f_o + f'(\lambda_j)}{f_o + f'(\lambda_o)}. \tag{14.3.5.7}$$

This set of F_o, $F_o + \Delta^{\rm ISO}_{\lambda_o}$ and $\Delta^{\rm ANO}_{\lambda_j}$ can then be used just as F_P, F_{PH} and $\Delta^{\rm ANO}$ are used in the SIRAS (single isomorphous replacement with anomalous scattering) method.

The algorithm described above is implemented in the program segment *MADMRG* as part of *Solve* (Terwilliger, 1994*b*). In most cases, there are more than one pair of X-ray wavelengths corresponding to a particular reflection. The estimates from each pair of wavelengths are averaged, using weighting factors based on the uncertainties in each estimate. Data from various pairs of X-ray wavelengths and from various Bijvoet pairs can have very different weights in their contributions to the total. This can be understood by noting that pairs of wavelengths that yield a large value of the denominator in equation (14.3.5.6) (*i.e.*, those that differ considerably in dispersive contributions) would yield relatively accurate estimates of $\Delta^{\rm ISO}_{\lambda_o}$. In the same way, Bijvoet differences measured at the wavelength with the largest value of f'' will contribute the most to estimates of $\Delta^{\rm ANO}_{\lambda_j}$.

The standard wavelength choice in this analysis is arbitrary, because values at any wavelength can be converted to values at any other wavelength. The standard wavelength does not even have to be one of the wavelengths in the experiment, though it is convenient to choose one of them.

14.3.6. Scoring of trial heavy-atom solutions

Scoring of potential heavy-atom solutions is an essential part of the *Solve* algorithm because it allows ranking of solutions and appropriate decision making. *Solve* scores trial heavy-atom solutions (or anomalously scattering atom solutions) using four criteria: agreeement with the Patterson function, cross-validation of heavy-atom sites, figure of merit, and non-randomness of the electron-density map. The scores for each criterion are normalized to those for a group of starting solutions (most of which are incorrect) to obtain Z scores. The total score for a solution is the sum of its Z scores after correction for anomalously high scores in any category.

The first criterion used by *Solve* for evaluating a trial heavy-atom solution is the agreement between calculated and observed Patterson functions. Comparisons of this type have always been important in the MIR and MAD methods (Blundell & Johnson, 1976). The score for Patterson-function agreement is the average value of the Patterson function at predicted locations of peaks, after multiplication by a weighting factor based on the number of heavy-atom sites in the trial solution. The weighting factor (Terwilliger & Berendzen, 1999*b*) is adjusted so that if two solutions have the same mean value at predicted Patterson peaks, the one with the larger numbers of sites receives the higher score. Typically the weighting factor is approximately given by $(N)^{1/2}$, where there are N sites in the solution.

In some cases, predicted Patterson vectors fall on high peaks that are not related to the heavy-atom solution. To exclude these contributions, occupancies of each heavy-atom site are refined so that the predicted peak heights approximately match the observed peak heights at the predicted interatomic positions. Then all peaks with heights more than 1σ higher than their predicted values are truncated at this height. The average values are further corrected for instances where more than one predicted Patterson vector falls on the same location by scaling that peak height by the fraction of predicted vectors that are unique.

A 'cross-validation' difference Fourier analysis is the basis of the second criterion used to evaluate heavy-atom solutions. One at a time, each site in a solution (and any equivalent sites in other derivatives for MIR solutions) is omitted from the heavy-atom model and phases are recalculated. These phases are used in a difference Fourier analysis and the peak height at the location of the omitted site is noted. A similar analysis where a derivative is omitted from phasing and all other derivatives are used to phase a difference Fourier has been used for many years (Dickerson *et al.*, 1961). The score for cross-validation difference Fouriers is the average peak height, after weighting by the same factor used in the difference Patterson analysis.

The mean figure of merit of phasing (m) (Blundell & Johnson, 1976) can be a remarkably useful measure of the quality of phasing despite its susceptibility to systematic error (Terwilliger & Berendzen, 1999*b*). The overall figure of merit is essentially a measure of the internal consistency of the heavy-atom solution and the data, and is used as the third criterion for solution quality in *Solve*. As heavy-atom refinement in *Solve* is carried out using origin-removed Patterson refinement (Terwilliger & Eisenberg, 1983), occupancies of heavy-atom sites are relatively unbiased. This minimizes the problem of high occupancies leading to inflated figures of merit. Additionally, using a single procedure for phasing allows comparison between solutions. The score based on figure of merit is simply the unweighted mean for all reflections included in phasing.

The most important criterion used by a crystallographer in evaluating the quality of a heavy-atom solution is the interpretability of the resulting electron-density map. Although a full implementation of such a criterion is difficult, it is quite straightforward to evaluate instead whether the electron-density map has features that are expected for a crystal of a macromolecule. A number of features of electron-density maps could be used for this purpose, including the connectivity of electron density in the maps (Baker *et al.*, 1993), the presence of clearly defined regions of protein and solvent (Wang, 1985; Podjarny *et al.*, 1987; Zhang & Main, 1990; Xiang *et al.*, 1993; Abrahams *et al.*, 1994; Terwilliger & Berendzen, 1999*a*,*c*), and histogram matching of electron densities (Zhang & Main, 1990; Goldstein & Zhang, 1998). We have used the identification of solvent and protein regions as the measure of map quality in *Solve*. This requires that there be both solvent and protein regions in the electron-density map, but for most macromolecular structures the fraction of the unit cell that is occupied by the macromolecule is in the suitable range of 30–70%. The criterion used in scoring by *Solve* is based on the connectivity of the solvent and protein regions (Terwilliger & Berendzen, 1999*c*). The unit cell is divided into boxes approximately twice the resolution of the map on a side, and within each box the r.m.s. electron density is calculated, without including the F_{000} term in the Fourier synthesis. For boxes within the protein region, this r.m.s. electron density will typically be high (as there are some points where atoms are located and other points between atoms), while for those in the solvent region it will be low (as the electron density is fairly uniform). The score based on the connectivity of the protein and solvent regions is simply the correlation coefficient of this r.m.s. electron density for adjacent boxes. If there is a large contiguous protein region and a large contiguous solvent region, then adjacent boxes will have highly correlated values of their r.m.s. electron densities. If the electron density is random, there will be little or no correlation. In practice, for a very good electron-density map, this correlation of local r.m.s. electron density may be as high as 0.5 or 0.6.

14.3.7. Automated MIR and MAD structure determination

The four-point scoring scheme described above provides the foundation for automated structure solution. To make it practical, the conversion of MAD data to a pseudo-SIRAS form and the use of rapid origin-removed Patterson-based heavy-atom refinement has been nearly essential. The remainder of the *Solve* algorithm for automated structure solution is largely a standardized form of local scaling, an integrated set of routines to carry out all of the calculations required for heavy-atom searching, refinement and phasing, and routines to keep track of the lists of current solutions being examined and past solutions that have already been tested.

Scaling of data in the *Solve* algorithm is done by a local scaling procedure (Matthews & Czerwinski, 1975). Systematic errors are minimized by scaling F^+ and F^-, native and derivative, and wavelengths of MAD data in very similar ways and by keeping different data sets separate until the end of scaling. The scaling procedure is optimized for cases where the data are collected in a systematic fashion. For both MIR and MAD data, the overall procedure is to construct a reference data set that is as complete as possible and that contains information either from a native data set (for MIR) or for all wavelengths (for MAD data). This reference data set is constructed for just the asymmetric unit of data and is essentially the average of all measurements obtained for each reflection. The reference data set is then expanded to the entire reciprocal lattice and used as the basis for local scaling of each individual data set [see Terwilliger & Berendzen (1999*b*) for additional details].

Once MIR data have been scaled, or MAD data have been scaled and converted to a pseudo-SIRAS form, difference Patterson functions are used to identify plausible one-site or two-site heavy-atom solutions. For MIR data, difference Patterson functions are calculated for each derivative. For MAD data, anomalous and dispersive differences are combined to yield a Bayesian estimate of the Patterson function for the anomalously scattering atoms (Terwilliger, 1994*a*). An automated search of the Patterson function is then used to find a large number (typically 30) of potential single-site and two-site solutions. In principle, Patterson methods could be used to solve the complete heavy-atom substructure, but the approach used in *Solve* is to find just the first one or two heavy-atom sites in this way and to find all others by difference Fourier analysis. This initial set of one-site and two-site solutions becomes the initial list of potential solutions ('seeds') for automated structure solution. Once each of the potential seeds is scored and ranked, the top seeds (typically five) are selected as independent starting points for the search for heavy-atom solutions.

For each starting solution (seed), the main cycle in the automated structure-solution algorithm used by *Solve* consists of two basic steps. The first is to refine heavy-atom parameters and rank all existing solutions generated so far from this seed based on the four criteria discussed above. The second is to take the highest-ranking solution that has not yet been exhaustively analysed and use it in an attempt to generate a more complete solution. Generation of new solutions is carried out in three ways: by deletion of sites, by addition of sites from difference Fouriers and by inversion. A partial solution is considered to have been exhaustively analysed when all single-site deletions have been considered, when no more peaks in a difference Fourier can be found that improve upon it, and when inversion does not improve it, or when the maximum number of sites input by the user has been reached. In each case, new solutions generated in these three ways are refined, scored and ranked, and the cycle is continued until all the top solutions have been fully analysed and no new solutions are found. Throughout this process, a tally of the solutions that have already been considered is kept, and any time a solution is a duplicate of a previously examined solution it is dropped.

In some cases, one very clear solution appears early in the structure-solution process, while in others, there are several solutions that have similar scores at early (and sometimes even late) stages of structure solution. In cases where no one solution is much better than the others, all the seeds are exhaustively analysed. On the other hand, if a very promising solution emerges from one seed, then the search is narrowed to focus on that seed, deletions are not carried out until the end of the analysis, and many peaks from the difference Fourier analysis are added at a time so as to build up the solution as quickly as possible. Once the expected number of heavy-atom sites are found, then each site is deleted in turn to see if the solution can be further improved. If this occurs, then the new solutions are analysed in the same way by addition and deletion of sites and by inversion until no improvement is obtained.

At the conclusion of the *Solve* algorithm, an electron-density map and phases for the top solution are reported in a form that is compatible with the *CCP4* suite (Collaborative Computational Project, Number 4, 1994). Additionally, command files that can be modified to look for additional heavy-atom sites or to construct other electron-density maps are produced. If more than one possible solution is found, the heavy-atom sites and phasing statistics for all of them are reported.

14.3.8. Generation of model X-ray data sets

An important feature of *Solve* is the inclusion of modules for the generation of model data. *Solve* can construct model raw X-ray data for either MIR or MAD cases. The macromolecular structure can be defined by a file in PDB format (Bernstein *et al.*, 1977) with heavy-atom parameters defined by the user. Any degree of 'experimental' uncertainty in measurement of intensities can be included, and limited non-isomorphism for MIR data in which cell dimensions differ for native and any of the derivative data sets (but in which the macromolecular structure is identical) can be included. This automatic generation of model data is very useful in evaluating what can and what cannot be solved. Once a data set has been generated, the *Solve* algorithm can be used to attempt to solve it. *Solve* generates a model electron-density map based on the input coordinates, and during the structure-solution process all maps calculated with trial solutions can be compared to the model map. In many cases, heavy-atom solutions can be related to different origins (and to different handedness as well). The origin shift is identified by *Solve* by finding the shift that best maps the trial solution onto the (known) correct solution.

14.3.9. Conclusions

The *Solve* algorithm is very useful for solving macromolecular structures by the MIR and MAD methods. It has been used to solve MAD structures with as many as 56 selenium atoms in the asymmetric unit (W. Smith & C. Janson, personal communication). From the user's point of view, the algorithm is very simple. Only a few input parameters are needed in most cases, and the *Solve* algorithm carries out the entire process automatically. In principle, the procedure can be very thorough as well, so that many trial starting solutions can be examined and difficult heavy-atom structures can be found. Additionally, for the most difficult structure-solution cases, the failure to find a solution can be useful in confirming that additional information is needed.

14.3.10. Software availability

The *Solve* software and complete documentation can be obtained from the web site http://solve.lanl.gov/.

TCT and JB gratefully acknowledge support from the National Institutes of Health and the US Department of Energy.

References

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Structure at 2.8-angstrom resolution of f1-ATPase from bovine heart-mitochondria. Nature (London)*, **370**, 621–628.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Uniqueness and the ab initio phase problem in macromolecular crystallography. Acta Cryst.* D**49**, 186–192.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *Protein data bank: computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, p. 368. New York: Academic Press.

Chang, G. & Lewis, M. (1994). *Using genetic algorithms for solving heavy-atom sites. Acta Cryst.* D**50**, 667–674.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *The crystal structure of myoglobin: phase determination to a resolution of 2 Å by the method of isomorphous replacement. Acta Cryst.* **14**, 1188–1195.

Goldstein, A. & Zhang, K. Y. J. (1998). *The two-dimensional histogram as a constraint for protein phase improvement. Acta Cryst.* D**54**, 1230–1244.

La Fortelle, E. de & Bricogne, G. (1997). *Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol.* **276**, 472–494.

Matthews, B. W. & Czerwinski, E. W. (1975). *Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta Cryst.* A**31**, 480–487.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB: crystal structure determination via shake-and-bake. J. Appl. Cryst.* **27**, 613–621.

Podjarny, A. D., Bhat, T. N. & Zwick, M. (1987). *Improving crystallographic macromolecular images: the real-space approach. Annu. Rev. Biophys. Biophys. Chem.* **16**, 351–373.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures. Acta Cryst.* A**46**, 467–473.

Terwilliger, T. C. (1994*a*). *MAD phasing: Bayesian estimates of F_A. Acta Cryst.* D**50**, 11–16.

Terwilliger, T. C. (1994*b*). *MAD phasing: treatment of dispersive differences as isomorphous replacement information. Acta Cryst.* D**50**, 17–23.

Terwilliger, T. C. & Berendzen, J. (1996). *Correlated phasing of multiple isomorphous replacement data. Acta Cryst.* D**52**, 749–757.

Terwilliger, T. C. & Berendzen, J. (1997). *Bayesian correlated MAD phasing. Acta Cryst.* D**53**, 571–579.

Terwilliger, T. C. & Berendzen, J. (1999*a*). *Discrimination of solvent from protein regions in native Fouriers as a means of evaluating heavy-atom solutions in the MIR and MAD methods. Acta Cryst.* D**55**, 501–505.

Terwilliger, T. C. & Berendzen, J. (1999*b*). *Automated MIR and MAD structure solution. Acta Cryst.* D**55**, 849–861.

Terwilliger, T. C. & Berendzen, J. (1999*c*). *Evaluation of macromolecular electron-density map quality using the correlation of local r.m.s. density. Acta Cryst.* D**55**, 1872–1877.

Terwilliger, T. C. & Eisenberg, D. (1983). *Unbiased three-dimensional refinement of heavy-atom parameters by correlation of origin-removed Patterson functions. Acta Cryst.* A**39**, 813–817.

Terwilliger, T. C. & Eisenberg, D. (1987). *Isomorphous replacement: effects of errors on the phase probability distribution. Acta Cryst.* A**43**, 6–13.

Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Generalized method of determining heavy-atom positions using the difference Patterson function. Acta Cryst.* A**43**, 1–5.

Vagin, A. & Teplyakov, A. (1998). *A translation-function approach for heavy-atom location in macromolecular crystallography. Acta Cryst.* D**54**, 400–402.

Wang, B.-C. (1985). *Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol.* **115**, 90–112.

Xiang, S., Carter, C. W. Jr, Bricogne, G. & Gilmore, C. J. (1993). *Entropy maximization constrained by solvent flatness: a new method for macromolecular phase extension and map improvement. Acta Cryst.* D**49**, 193–212.

Zhang, K. Y. J. & Main, P. (1990). *The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. Acta Cryst.* A**46**, 377–381.

International Tables for Crystallography (2012). Vol. F, Chapter 15.1, pp. 385–400.

PART 15. DENSITY MODIFICATION AND PHASE COMBINATION

Chapter 15.1. Phase improvement by iterative density modification

K. Y. J. Zhang, K. D. Cowtan and P. Main

15.1.1. Introduction

Density modification is a technique for improving the quality of an approximate electron-density map based on some conserved features of the correct electron-density map. These conserved features are independent of the unknown fine detail of the structural conformation. They are often expressed as constraints on the electron density in various forms, either in real or reciprocal space. Since the structure-factor amplitudes are known, these constraints restrict the values of phases and can therefore be used for phase improvement.

The structure-factor amplitudes and phases are independent of each other if we know nothing about the electron density. Therefore, the phases are indeterminable given only the amplitudes (Baker, Krukowski & Agard, 1993). The information about the electron density provides the missing link between structure-factor amplitudes and phases. It is only through the knowledge of the chemical or physical properties of the electron density that the phases can be retrieved. Density modification is usually the most straightforward application of the constraints on electron density. However, this is only a matter of convenience in implementation. Sometimes the constraints can be more readily implemented in reciprocal space on structure factors.

Density-modification methods are usually implemented as an iterative procedure that alternates between density modification in real space and phase combination in reciprocal space. This paradigm was first proposed by Hoppe & Gassmann (1968) in their 'phase correction' method. This approach takes advantage of the particular properties of the constraints and uses them in a way that is most convenient to implement.

Density-modification methods usually require an initial map with substantial phase information. In most cases, these phases are obtained from multiple isomorphous replacement (MIR) or multiwavelength anomalous dispersion (MAD), but it is also possible to improve maps from other sources, such as molecular replacement. The amount of information in the initial map is dependent on phase accuracy, data resolution and completeness. As more powerful constraints are incorporated, the density modification can be initiated from lower-resolution maps with less accurate phases. *Ab initio* phasing would be achieved if a density-modification method could start from a map generated from random phases. Therefore, density modification can potentially lead to *ab initio* phasing methods, although it does not seek direct solution to the phase problem as its immediate goal.

There are two major components in a density-modification procedure. One is the type of electron-density constraints. The other is the way the constraints are exploited. These two components combined determine the phasing power of the procedure. In this chapter, we will review various electron-density constraints and the way they are exploited for phase improvement.

15.1.2. Density-modification methods

The aim of density-modification calculations is to obtain new or improved phase estimates for observed structure-factor amplitudes. Often, this includes calculation of phases for previously unphased reflections, for example, in the case of phase extension. The calculation of weights, which indicate the degree of confidence in the new phase estimates, is also an important part of the calculation. Improved phase estimates are obtained by bringing the initial phase estimates into consistency with additional sources of structural information.

One difficulty in combining information from various sources is that the amplitudes and phases are represented in reciprocal space and include good estimates of error, whereas the other constraints are in real space and in general, represent expectations about the structure which may be hard to quantify. As a result, the method that has been adopted is iterative and divided into real- and reciprocal-space steps. A weighted map is calculated and used as a basis for applying all the real-space modifications. The modified map is then back-transformed to produce a set of amplitudes and phases. The agreement between the observed amplitudes and the amplitudes calculated from the modified map is then used to estimate weights for the modified phases, which are used to combine the modified phases with experimental phases to produce new phases. This process is shown diagrammatically in Fig. 15.1.2.1.

A broad range of techniques have been applied to electron-density maps to impose chemical or physical information. Some sources of information used in density modification are summarized in Table 15.1.2.1. The list included here is not exhaustive, but covers the most widely used methods. Here, we describe some of the constraints and the techniques through which these constraints are implemented for phase improvement.

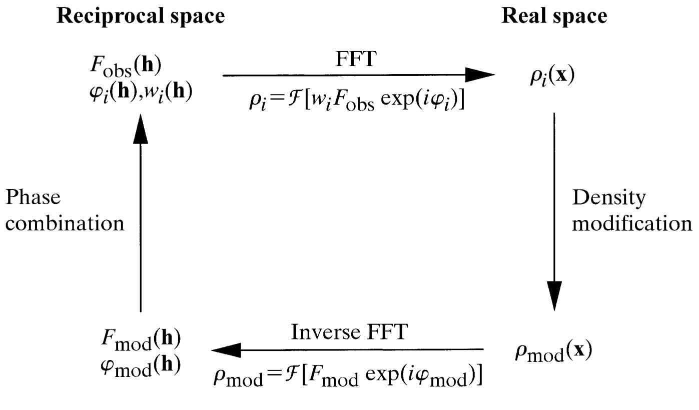

Reciprocal space
Real space
$F_{\rm obs}(\mathbf{h})$
$\varphi_i(\mathbf{h}), w_i(\mathbf{h})$
FFT
$\rho_i = \mathcal{F}[w_i F_{\rm obs} \exp(i\varphi_i)]$
$\rho_i(\mathbf{x})$
Phase combination
Density modification
$F_{\rm mod}(\mathbf{h})$
$\varphi_{\rm mod}(\mathbf{h})$
Inverse FFT
$\rho_{\rm mod} = \mathcal{F}[F_{\rm mod} \exp(i\varphi_{\rm mod})]$
$\rho_{\rm mod}(\mathbf{x})$

Figure 15.1.2.1
Density-modification calculation showing iterative application of real-space and reciprocal-space constraints.

Copyright © 2012 International Union of Crystallography

Table 15.1.2.1
Constraints used in density modification

Constraints	Use	Effectiveness and limitation
(1) Solvent flatness	Solvent flattening	Works best at medium resolution. Relatively resolution insensitive. Good for phase refinement. Weak on phase extension.
(2) Ideal electron-density distribution	Histogram matching	Works at a wide range of resolutions. More effective at higher resolution. Very effective for phase extension.
(3) Equal molecules	Molecular averaging	Works better at low to medium resolution. Its phasing power increases with the number of molecules in the asymmetric unit.
(4) Protein backbone connectivity	Skeletonization	Requires near atomic resolution to work.
(5) Local shape of electron density	Sayre's equation	The equation is exact at atomic resolution. It can be used at non-atomic resolution by choosing an appropriate shape function. Its phasing power increases quickly with resolution. Very powerful for phase extension.
(6) Atomicity	Atomization	If the initial map is good enough, iteration could lead to a final model.
(7) Structure-factor amplitudes	Sim weighting	Can be used to estimate the reliability of the calculated phases after density modification. It assumes the random distribution of errors that caused the discrepancy between the calculated and observed structure-factor amplitudes.
(8) Experimental phases	Phase combination	This can be used to filter out the incorrect component of the estimated phases. Most phase-combination procedures assume independence between the calculated and observed phases.

15.1.2.1. Solvent flattening

Solvent flattening exploits the fact that the electron density in the solvent region is flat at medium resolution, owing to the high thermal motion and disorder of solvent molecules. The flattening of the solvent region suppresses noise in the map and therefore improves phases.

15.1.2.1.1. Introduction

Biological molecules are typically irregular in shape, often taking roughly globular forms. When they are packed regularly to form a crystal lattice, there are gaps left between them, and these spaces are filled with the solvent in which the crystallization was performed. This solvent is a disordered liquid, and thus the arrangement of atoms in the solvent regions varies between unit cells, except in those small regions near the surface of the protein. The X-ray image forms an average of electron density over many cells, so the electron density over much of the solvent region appears to be constant to a good approximation.

The existence of a flat solvent region in a crystal places strong constraints on the structure-factor phases. The constraint of solvent flatness is implemented by identifying the molecular boundaries and replacing the densities in the solvent region by their mean density value.

When solving a structure, the contents of the unit cell are usually known, and so an estimate can be formed of how much of the cell volume is taken up by solvent (Matthews, 1968). If the solvent region can be located in the cell, then we can improve an electron-density map by setting the electron density in this region to the expected constant solvent density. Once the resulting modified phases are combined with the experimental data, an improvement can often be seen in the protein regions of the map (Bricogne, 1974).

The solvent region of a unit cell may usually be determined even from a poor MIR map using the following features:

(1) The mean electron density in the solvent region should be lower than that in the protein region. Note that this information will come from the low-resolution data, which dictate long-range density variations over the unit cell.
(2) The variation in density in the flat solvent region should be much smaller than that in the ordered protein region containing isolated clumps of density. The 'peakiness' of the protein region comes from the high-resolution data.

A good method for locating the solvent region therefore takes into account information from both low- and high-resolution structure factors. Many methods have been proposed to locate the protein–solvent boundary. The first of these were the visual identification methods. The boundary was identified by digitizing a mini-map with the aid of a graphic tablet (Hendrickson *et al.*, 1975; Schevitz *et al.*, 1981). The hand-digitizing procedure was very time consuming and prone to subjective judgmental errors. Nevertheless, these methods demonstrated the potential of solvent flattening and stimulated further improvement on boundary-identification methods. An automated method using a linked, high-density approach was first proposed by Bhat & Blow (1982). Based on the fact that the densities are generally higher in the protein region than in the solvent region, they defined the molecular boundary by locating the protein as a region of linked, high-density points.

Convolution techniques were subsequently adopted as an efficient method of molecular-boundary identification. Reynolds *et al.* (1985) proposed a high mean absolute density value approach. The electron density within the protein region was expected to have greater excursions from the mean density value than the solvent region, which is relatively featureless. The molecular boundary was located based on the value of a smoothed ‘modulus’ electron density, which is the sum of the absolute values of all density points within a small box.

15.1.2.1.2. The automated convolution method for molecular-boundary identification

Wang (1985) suggested an automated convolution method for identifying the solvent region which has achieved widespread use. His method involved first calculating a truncated map:

$$\rho_{\rm trunc}(\mathbf{x}) = \begin{cases} \rho\mathbf{x}, & \rho(\mathbf{x}) > \rho_{\rm solv} \\ 0, & \rho(\mathbf{x}) < \rho_{\rm solv} \end{cases}. \qquad (15.1.2.1)$$

The electron density is simply truncated at the expected solvent value, $\rho_{\rm solv}$; however, since the variations in density in the protein region are much larger than the variations in the solvent region, it is generally only the protein region which will be affected. Thus, the mean density over the protein region is increased. Similar results may be obtained using the mean-squared difference of the density from the expected solvent value.

A smoothed map is then formed by calculating at each point in the map the mean density over a surrounding sphere of radius R. This operation can be written as a convolution of the truncated map, $\rho_{\rm trunc}$, with a spherical weighting function, $w(\mathbf{r})$,

$$\rho_{\rm ave}(\mathbf{x}) = \sum_{\mathbf{r}} {\rm w}(\mathbf{r})\rho_{\rm trunc}(\mathbf{x}-\mathbf{r}), \qquad (15.1.2.2)$$

where

$$w(\mathbf{r}) = \begin{cases} 1 - |\mathbf{r}|/R, & |\mathbf{r}| < R \\ 0, & |\mathbf{r}| > R \end{cases}. \qquad (15.1.2.3)$$

Leslie (1987) noted that the convolution operation required in equation (15.1.2.2) can be very efficiently performed in reciprocal space using fast Fourier transforms (FFTs),

$$\rho_{\rm ave}(\mathbf{x}) = \mathcal{F}^{-1}\{\mathcal{F}[\rho_{\rm trunc}(\mathbf{x})]\mathcal{F}[w(\mathbf{r})]\}, \qquad (15.1.2.4)$$

where $\mathcal{F}$ denotes a Fourier transform, and $\mathcal{F}^{-1}$ represents an inverse Fourier transform.

The Fourier transform of the truncated density can be readily calculated using FFTs. The Fourier transform of the weighting function can be calculated analytically by

$$g(s) = \mathcal{F}[w(\mathbf{r})] = \frac{3[\sin(2\pi Rs) - 2\pi Rs\cos(2\pi Rs)]}{(2\pi Rs)^3} - \frac{3\{4\pi Rs\sin(2\pi Rs) - [(2\pi Rs)^2 - 2]\cos(2\pi Rs) - 2\}}{(2\pi Rs)^4}, \qquad (15.1.2.5)$$

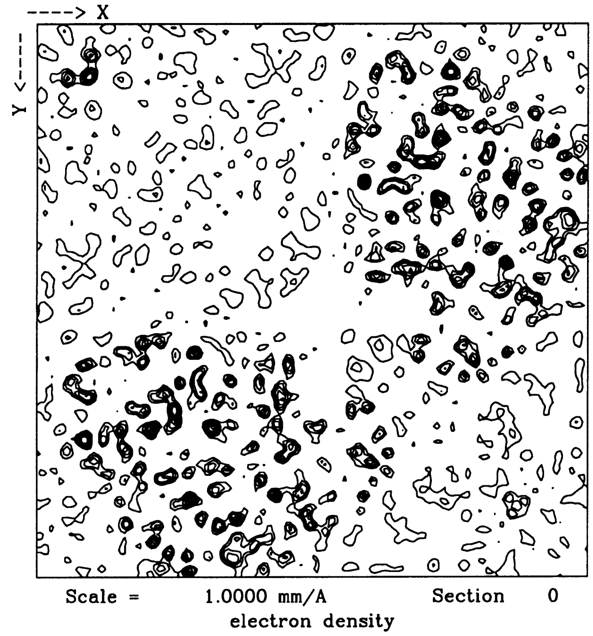
where

$$s = 2\sin\theta/\lambda.$$

Therefore, the averaging of the truncated electron density by a spherical weighting function can be achieved by two FFTs. This greatly reduced the time required for calculating the averaged density. Other weighting functions may be implemented by the same approach.

A cutoff value, $\rho_{\rm cut}$, is then calculated, which divides the unit cell into two portions occupying the correct volumes for the protein and solvent regions. All points in the map where $\rho_{\rm ave}(\mathbf{x}) < \rho_{\rm cut}$ can then be assumed to be in the solvent region. A typical mask obtained from an MIR map by this means, and the modified map, are shown in Fig. 15.1.2.2.

The radius of the sphere, R, used in equation (15.1.2.3) for the averaging of electron densities is generally around 8 Å. The molecular envelope derived from such an averaged map tends to lose details of the protein molecular surface. Paradoxically, a large averaging sphere is required for the identification of the protein–solvent boundary based on the difference between the mean density of the protein and solvent, which is very small and can only be distinguished when a sufficiently large area of the map is averaged. Abrahams & Leslie (1996) proposed an alternative method of molecular-boundary identification that uses the standard deviation of the electron density within a given radius relative to the overall mean at every grid point of a map. The local-standard-deviation map is the square root of a convolution of a sphere and the squared map, which can be calculated in reciprocal space in a similar way to the procedure described in equations (15.1.2.4) and (15.1.2.5) as proposed by Leslie (1987). By integrating the histogram of the local-standard-deviation map, the cutoff value of the local standard deviation corresponding to the solvent fraction can be calculated. Using this procedure, a molecular envelope that contains more details of the protein molecular surface can be obtained, since the radius of the averaging sphere can be as low as 4 Å (Abrahams & Leslie, 1996).



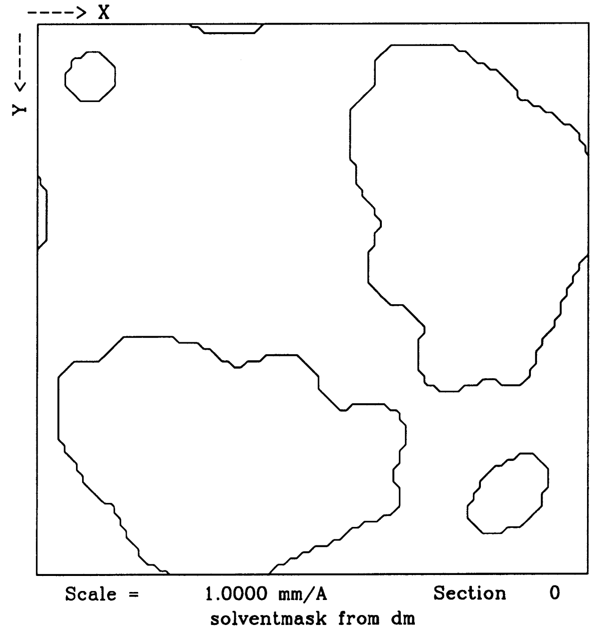


Figure 15.1.2.2
Solvent mask determined from a map by Wang’s method.

15.1.2.1.3. *The solvent-flattening procedure*

Once the envelope has been determined, solvent flattening is performed by simply setting the density in the solvent region to the expected value, $\rho_{\rm solv}$:

$$\rho_{\rm mod}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\rm ave}(\mathbf{x}) > \rho_{\rm cut} \\ \rho_{\rm solv}, & \rho_{\rm ave}(\mathbf{x}) < \rho_{\rm cut} \end{cases}. \qquad (15.1.2.6)$$

If the electron density has not been calculated on an absolute scale, the solvent density may be set to its mean value.

A related method is solvent flipping, developed by Abrahams & Leslie (1996). In this approach, the flattening operation is modified by the introduction of a relaxation factor, γ, where γ is positive, effectively 'flipping' the density in the solvent region.

$$\rho_{\rm mod}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\rm ave}(\mathbf{x}) > \rho_{\rm cut} \\ \rho_{\rm solv} - [\gamma/(1-\gamma)][\rho(\mathbf{x}) - \rho_{\rm solv}], & \rho_{\rm ave}(\mathbf{x}) < \rho_{\rm cut} \end{cases}. \qquad (15.1.2.7)$$

The effect of this modification is to correct for the problem of independence in phase combination and is discussed in Section 15.1.4.3.

15.1.2.2. *Histogram matching*

Histogram matching seeks to bring the distribution of electron-density values of a map to that of an ideal map. The density histogram of a map is the probability distribution of electron-density values. It provides a global description of the appearance of the map, and all spatial information is discarded. The comparison of the histogram for a given map with that expected for an ideal map can serve as a measure of quality. Furthermore, the initial map can be improved by adjusting density values in a systematic way to make its histogram match the ideal histogram.

15.1.2.2.1. *Introduction*

Histogram matching is a standard technique in image processing. It is aimed at bringing the density distribution of an image to an ideal distribution, thereby improving the image quality. The first attempt at modifying the electron-density distribution was that by Hoppe & Gassman (1968), who proposed the '3–2' rule. The electron density was first normalized to a maximum of 1 and modified by imposing positivity. Subsequently, the electron density was modified by $\rho_{\rm mod} = 3\rho^2 - 2\rho^3$. Podjarny & Yonath (1977) used the skewness of the density histogram as a measure of quality of the modified map. Harrison (1988) used a Gaussian function as the ideal histogram in his histogram-specification method for protein phase refinement and extension. The choice of the Gaussian function as the ideal electron-density distribution was based on theoretical arguments instead of experimental evaluation. The Gaussian function was also made independent of resolution. Lunin (1988) used the electron-density distribution to retrieve the values of low-angle structure factors whose amplitudes had not been measured during an X-ray experiment. The electron-density distribution was thought to be structure specific and was derived from a homologous structure. Moreover, the histogram was derived from the entire unit cell, including both the protein and the solvent. Zhang & Main (1988) systematically examined the electron-density histogram of several proteins and found that the ideal density histogram is dependent on resolution, the overall temperature factor and the phase error. It is, however, independent of structural conformation. The sensitivity to phase error suggests that the density histogram could be used for phase improvement. The structural conformation independence made it possible to predict the ideal histogram for unknown structures.

15.1.2.2.2. *The prediction of the ideal histogram*

Polypeptide structures in particular, and biological macromolecules in general, display a broadly similar atomic composition, and the way in which these atoms bond together is also conserved across a wide range of structures. These similarities between different protein structures can be used to predict the ideal histogram even when positional information for individual atoms is not available in a map. If the positional information is removed from an electron-density map, then what remains is an unlabelled list of density values. This list is the histogram of the electron-density distribution, which is independent of the relative disposition of these densities. The shape of the histogram is primarily based on the presence of atoms and their characteristic distances from each other. This is true for all polypeptide structures.

The frequency distribution, $P(\rho)$, of electron-density values in a map can be constructed by sampling the map and counting the density values in different ranges. In practice, once the electron-density map has been sampled on a discrete grid, this frequency distribution becomes a histogram, but for convenience, it is treated here as a continuous distribution.

At resolutions of better than 6.0 Å and after exclusion of the solvent region, the frequency distribution of electron-density values for protein density over a wide range of proteins varies only with resolution and overall temperature factor to a good approximation. If the overall temperature factor is artificially adjusted, for example, by sharpening to $B_{\rm overall} = 0$, then the frequency distributions may be treated as a function of resolution only. Therefore, once a good approximation to the molecular envelope is known, the frequency distribution of electron densities in the protein region as a function of resolution may be assumed to be known. Therefore, the ideal density histogram for an unknown map at a given resolution can be taken from any known structure at the same resolution (Zhang & Main, 1988, 1990*a*).

The ideal electron-density histogram can also be predicted by an analytical formula (Lunin & Skovoroda, 1991; Main, 1990*a*). The method adopted by Main (1990*a*) represents the density histogram by components that correspond to three types of electron density in the map. The first component is the region of overlapping densities, which can be represented by a randomly distributed background noise. The second component is the region of partially overlapping densities. The third component is the region of non-overlapping atomic peaks, which can be represented by a Gaussian.

The histogram for the overlapping part of the density can be represented by a Gaussian distribution,

$$P_o(\rho) = N \exp\bigl[-(\rho - \overline{\rho})^2/2\sigma^2\bigr], \qquad (15.1.2.8)$$

where $\overline{\rho}$ is the mean density and σ is the standard deviation. The region of partially overlapping densities can be modelled by a cubic polynomial function,

$$P_{po}(\rho) = N\bigl(a\rho^3 + b\rho^2 + c\rho + d\bigr). \qquad (15.1.2.9)$$

The histogram for the non-overlapping part of the density can be derived analytically from a Gaussian atom,

$$P_{no}(\rho) = N(A/\rho)[\ln(\rho_0/\rho)]^{1/2}, \qquad (15.1.2.10)$$

where ρ_0 is the maximum density, N is a normalizing factor and A

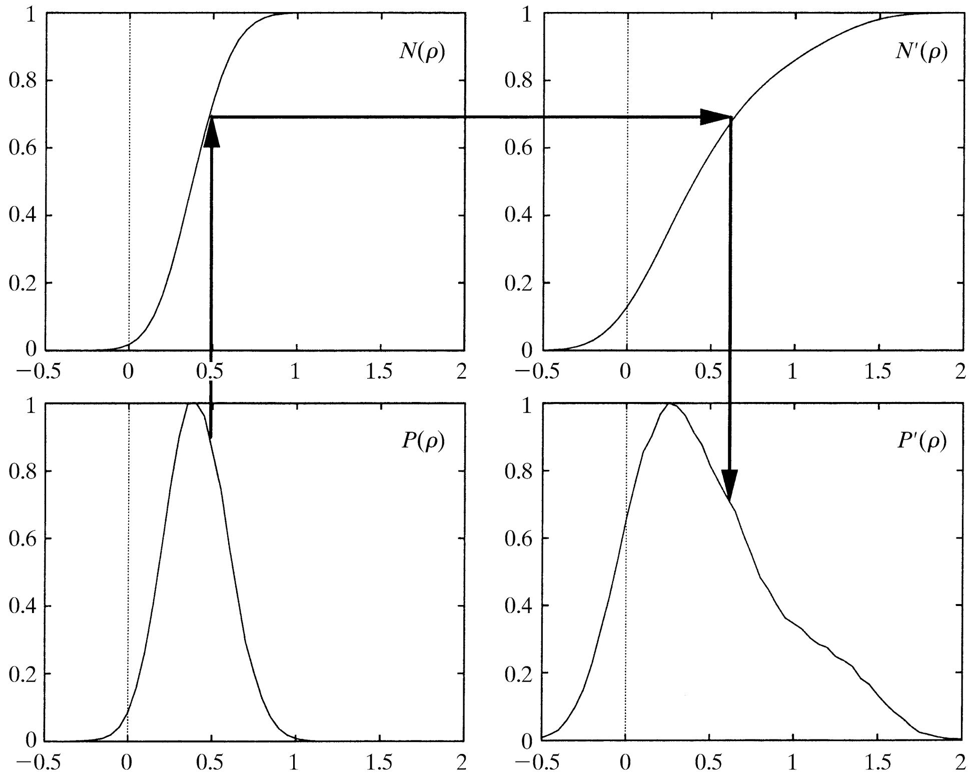


Figure 15.1.2.3
Transformation of density ρ to $\rho'_{\rm mod}$ by histogram matching.

is the relative weight of the terms between equation (15.1.2.8) and equation (15.1.2.10).

If we use two threshold values, ρ_1 and ρ_2, to divide the three density regions, the complete formula can be expressed as

$$P(\rho) = \begin{cases} N \exp\left[-(\rho - \overline{\rho})^2/2\sigma^2\right] & \text{for} \quad 2\rho \le \rho_2 \\ N(a\rho^3 + b\rho^2 + c\rho + d) & \text{for} \quad 2\rho_2 < \rho \le \rho_1 \\ N(A/\rho)[\ln(\rho_0/\rho)]^{1/2} & \text{for} \quad 2\rho_1 < \rho \le \rho_0. \end{cases} \tag{15.1.2.11}$$

The parameters a, b, c, d in the cubic polynomial are calculated by matching function values and gradients at ρ_1 and ρ_2. The parameters in the histogram formula, $\overline{\rho}$, σ, A, ρ_0, ρ_1, ρ_2, can be obtained from histograms of known structures.

15.1.2.2.3. The process of histogram matching

Zhang & Main (1990*a*) demonstrated that, at better than 4 Å resolution, the histogram for an MIR map is generally significantly different from the ideal distribution calculated from atomic coordinates. The obvious course is therefore to alter the map in such a way as to make its density histogram equal to the ideal distribution. Unfortunately, there are an infinite number of maps corresponding to any chosen density distribution, so we must choose a systematic method of altering the map.

The conventional method of performing such a modification is to retain the ordering of the density values in the map. The highest point in the original map will be the highest point in the modified map, the second highest points will correspond in the same way, and so on.

Mathematically, this transformation is represented as follows. Let $P(\rho)$ be the current density histogram and $P'(\rho)$ be the desired distribution, normalized such that their sums are equal to 1. The cumulative distribution functions, $N(\rho)$ and $N'(\rho)$, may then be calculated:

$$\begin{aligned} N(\rho) &= \int_{\rho_{\min}}^{\rho} P(\rho)\, {\rm d}\rho, \\ N'(\rho') &= \int_{\rho_{\min}}^{\rho'} P'(\rho)\, {\rm d}\rho. \end{aligned} \tag{15.1.2.12}$$

The cumulative distribution function of a variable transforms a value chosen from the distribution into a number between 0 and 1, representing the position of that value in an ordered list of values chosen from the distribution.

The transformation may, therefore, be performed in two stages. A density value is taken from the initial distribution and the cumulative distribution function of the initial distribution is applied to obtain the position of that value in the distribution. The inverse of the cumulative distribution function for the desired distribution is applied to this value to obtain the density value for the corresponding point in the desired distribution. Thus, given a density value, ρ, from the initial distribution, the modified value, ρ', is obtained by

$$\rho' = N'^{-1}[N(\rho)]. \tag{15.1.2.13}$$

The distribution of ρ' will then match the desired distribution after the above transformation. The transformation of an electron-density value by this method is illustrated in Fig. 15.1.2.3. The transformation in equation (15.1.2.13) can be achieved through a linear transform represented by

$$\rho'_i = a_i \rho_i + b_i, \tag{15.1.2.14}$$

where $i = \{1, \ldots, n\}$ and n is the number of density bins. The above linear transform is sufficient if the number of density bins is large enough. An n value of about 200 is usually quite satisfactory.

Various properties of the electron density are specified in the density histogram, such as the minimum, maximum and mean density, the density variance, and the entropy of the map. The mean density of the ideal map can be obtained by

$$\overline{\rho} = \int_{\rho_{\min}}^{\rho_{\max}} \rho P(\rho)\,\mathrm{d}\rho. \qquad (15.1.2.15)$$

The variance of the density in the ideal map can be obtained by

$$\sigma(\rho) = \left(\overline{\rho^2} - \overline{\rho}^2\right)^{1/2}, \qquad (15.1.2.16)$$

where

$$\overline{\rho^2} = \int_{\rho_{\min}}^{\rho_{\max}} \rho^2 P(\rho)\,\mathrm{d}\rho. \qquad (15.1.2.17)$$

Therefore, the process of histogram matching applies a minimum and a maximum value to the electron density, imposes the correct mean and variance, and defines the entropy of the new map. The order of electron-density values remains unchanged after histogram matching.

Histogram matching is complementary to solvent flattening since it is applied to the protein region of a map, whereas solvent flattening only operates on the solvent region of the map. The same envelope that was used for isolating the solvent region can be used to determine the protein region of the cell. An alternative approach is to define separate solvent and protein masks, with uncertain regions excluded from either mask and allowed to keep their unmodified values.

15.1.2.2.4. Scaling the observed structure-factor amplitudes according to the ideal density histogram

In the process of density modification, electron density or structure factors from different sources are compared and combined. It is, therefore, crucial to ensure that all the structure factors and maps are on the same scale. The observed structure factors can be put on the absolute scale by Wilson statistics (Wilson, 1949) using a scale and an overall temperature factor. This is accurate when atomic or near atomic resolution data are available. The scale and overall temperature factor obtained from Wilson statistics are less accurate when only medium- to low-resolution data are available. A more robust method of scaling non-atomic resolution data is through the density histogram (Cowtan & Main, 1993; Zhang, 1993).

The ideal density histogram defines the mean and variance of an electron density, as shown in equations (15.1.2.15) and (15.1.2.16). We can scale the observed structure-factor amplitudes to be consistent with the target histogram using the following formula, obtained from the structure-factor equation and Parseval's theorem. The mean density and the density variance of the observed map can be calculated as

$$\overline{\rho}' = (1/V)F(000), \qquad (15.1.2.18)$$

$$\sigma'(\rho) = (1/V)\left[\sum_{\mathbf{h}} |F(\mathbf{h})|^2\right]^{1/2}. \qquad (15.1.2.19)$$

The mean and variance of the electron-density map at the desired resolution are calculated using the target histogram, the mean value of the solvent density, $\overline{\rho}_{\rm solv}$, and the solvent volume of the cell, $V_{\rm solv}$. The $F(000)$ term can then be evaluated from equations (15.1.2.15) and (15.1.2.18):

$$F(000) = (V - V_{\rm solv})\overline{\rho} + V_{\rm solv}\overline{\rho}_{\rm solv}. \qquad (15.1.2.20)$$

The scale of the observed amplitudes can be obtained from equations (15.1.2.16) and (15.1.2.19),

$$F'(\mathbf{h}) = KF(\mathbf{h}), \qquad (15.1.2.21)$$

where

$$K = \left[(\overline{\rho^2} - \overline{\rho}^2)\right]^{1/2} \Big/ \left\{(1/V)\left[\sum_{\mathbf{h}} |F(\mathbf{h})|^2\right]^{1/2}\right\}. \qquad (15.1.2.22)$$

This method is adequate for scaling observed structure factors at any resolution.

15.1.2.3. Averaging

The averaging method enforces the equivalence of electron-density values between grid points in the map related by noncrystallographic symmetry. The averaging procedure can filter noise, correct systematic error and even determine the phases *ab initio* in favourable cases (Chapman *et al.*, 1992; Tsao *et al.*, 1992).

15.1.2.3.1. Introduction

Noncrystallographic symmetry (NCS) arises in crystals when there are two or more of the same molecules in one asymmetric unit. Such symmetries are local, since they only apply within a sub-region of a single unit cell. A fivefold axis, for example, must be noncrystallographic, since it is not possible to tessellate objects with fivefold symmetry. Since the symmetry does not map the crystal lattice back onto itself, the individual molecules that are related by the noncrystallographic symmetry will be in different environments; therefore, the symmetry relationships are only approximate.

Noncrystallographic symmetries provide phase information by the following means. Firstly, the related regions of the map may be averaged together, increasing the ratio of signal to noise in the map. Secondly, since the asymmetric unit must be proportionally larger to hold multiple copies of the molecule, the number of independent diffraction amplitudes available at any resolution is also proportionally larger. This redundancy in sampling the molecular transform leads to additional phase information which can be used for phase improvement.

15.1.2.3.2. The determination of noncrystallographic symmetry

The self-rotation symmetry is now routinely solved by the use of a Patterson rotation function (Rossmann & Blow, 1962). The translation symmetry can be determined by a translation function (Crowther & Blow, 1967) when a search model, either an approximate structure of the protein to be determined or the structure of a homologous protein, is available. The searches of the Patterson rotation and translation functions are achieved typically using fast automatic methods, such as *X-PLOR* (Brünger *et al.*, 1987) or *AMoRe* (Navaza, 1994). In cases where no search model is available or the Patterson translation function is unsolvable, either the whole electron-density map, or a region which is expected to contain a molecule, may be rotated using the rotation solution and used as a search model in a phased translation function (Read & Schierbeek, 1988).

Once the averaging operators are determined, the mask can be determined using the local density correlation function as developed by Vellieux *et al.* (1995). This is achieved by a systematic search for extended peaks in the local density correlation, which must be carried out over a volume of several unit cells in order to guarantee finding the whole molecule. The local correlation function distinguishes those volumes of crystal space which map onto similar density under transformation by the averaging operator. Thus, in the case of improper NCS, a local correlation mask will cover only one monomer. In the case of a proper symmetry, a local correlation mask will cover the whole complex (Fig. 15.1.2.4*a*,*b*).

Special cases arise when there are combinations of crystallographic and noncrystallographic symmetries, of proper and improper symmetries, or when a noncrystallographic symmetry element maps a cell edge onto itself. In the latter case, the volume of matching density is infinite, and arbitrary limits must be placed upon the mask along one crystal axis.

15.1.2.3.3. The refinement of noncrystallographic symmetry

The initial NCS operation obtained from rotation and translation functions or heavy-atom positions can be fine-tuned by a density-space *R*-factor search in the six-dimensional rotation and translation space. The density-space *R* factor is defined as

$$R = \sum_{\mathbf{r}} |\rho(\mathbf{r}) - \rho(\mathbf{r}')| \Big/ \sum_{\mathbf{r}} |\rho(\mathbf{r}) + \rho(\mathbf{r}')|, \qquad (15.1.2.23)$$

where $\mathbf{r} = \{xyz\}$ is the set of Cartesian coordinates, $\mathbf{r}' = \Omega\mathbf{r}$ is the NCS-related set of coordinates of $\mathbf{r}$ and Ω represents the NCS operator.

The six-dimensional search is very time consuming. The search rate can be increased by using only a representative subset of grid points. The NCS operation is systematically altered to find the lowest density-space *R* factor for the selected subset of grid points.

The solution of the NCS operation from the six-dimensional search can be further refined by the following least-squares procedure. If $\rho(\mathbf{r})$ is related to $\rho(\mathbf{r}')$ by the NCS operation, Ω,

$$\rho(\mathbf{r}') = \rho(\Omega\mathbf{r}). \qquad (15.1.2.24)$$

Here, Ω is a function of ω, $\Omega = f(\omega)$, where $\omega = \{\alpha, \beta, \gamma, t_x, t_y, t_z\}$ represents the rotation and translation components of the NCS operation. The solution to the NCS parameters, ω, can be obtained by minimizing the density residual between the NCS-related molecules,

$$\varepsilon(\mathbf{r}) = \rho(\mathbf{r}) - \rho(\Omega\mathbf{r}), \qquad (15.1.2.25)$$

using a least-squares formula of the form

$$\left(\frac{\partial\rho}{\partial\omega}\right)^T \left(\frac{\partial\rho}{\partial\omega}\right) \Delta\omega = \left(\frac{\partial\rho}{\partial\omega}\right)^T \varepsilon(\mathbf{r}), \qquad (15.1.2.26)$$

where $\Delta\omega$ is the shift to the NCS parameters. Here,

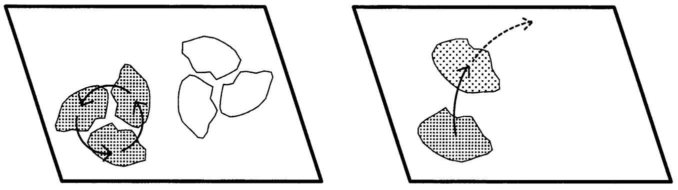
$$\frac{\partial\rho}{\partial\omega} = \frac{\partial\rho}{\partial\mathbf{r}}\frac{\partial\mathbf{r}}{\partial\omega}. \qquad (15.1.2.27)$$

The partial derivatives, $\partial\rho/\partial\mathbf{r} = \{\partial\rho/\partial x,\ \partial\rho/\partial y,\ \partial\rho/\partial z\}$, can be calculated by Fourier transforms,

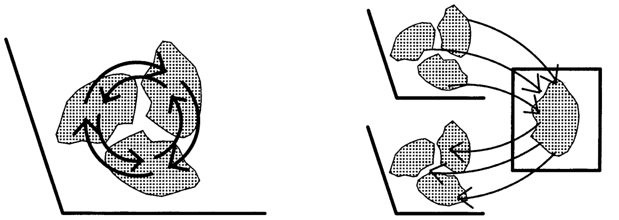
$$\begin{aligned}
\frac{\partial\rho}{\partial x} &= -\frac{2\pi i}{V}\sum_{hkl} hF_{hkl}\exp[-2\pi i(hx + ky + lz)] \\
\frac{\partial\rho}{\partial y} &= -\frac{2\pi i}{V}\sum_{hkl} kF_{hkl}\exp[-2\pi i(hx + ky + lz)] \qquad (5.1.2.28) \\
\frac{\partial\rho}{\partial z} &= -\frac{2\pi i}{V}\sum_{hkl} lF_{hkl}\exp[-2\pi i(hx + ky + lz)],
\end{aligned}$$

or more efficiently with a single Fourier transform by the use of spectral B-splines (Cowtan & Main, 1998). $\partial\mathbf{r}/\partial\omega$ is derived analytically based on the relationship between the Cartesian coordinates, *r*, and the rotational and translational coordinates of the NCS operation, ω,

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\alpha\cos\beta\cos\gamma - \sin\alpha\sin\gamma & -\cos\alpha\cos\beta\sin\gamma - \sin\alpha\sin\gamma & \cos\alpha\sin\beta \\ \sin\alpha\cos\beta\cos\gamma + \cos\alpha\sin\gamma & -\sin\alpha\cos\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta \\ -\sin\beta\cos\gamma & \sin\beta\sin\gamma & \cos\beta \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} t_x \\ t_y \\ t_z \end{pmatrix}. \qquad (15.1.2.29)$$



(*a*) Proper (closed) NCS (*b*) Improper (open) NCS



(*c*) One–stage averaging (*d*) Two–stage averaging

Figure 15.1.2.4
Types of noncrystallographic symmetry and averaging calculation.

15.1.2.3.4. The averaging of NCS-related molecules

Once the mask and matrices are determined, the electron-density map may be modified by averaging. This can be achieved in one or two stages: The density for each copy of the molecule in the asymmetric unit may be replaced by the averaged density from every copy; however, this becomes slow for high-order NCS (Fig. 15.1.2.4*c*). Alternatively, a single averaged copy of the molecule may be created in an artificial cell [referred to by Rossmann *et al.* (1992) as an *H*-cell], and then each copy of the molecule may be reconstructed in the asymmetric unit from this copy (Fig. 15.1.2.4*d*). This is more efficient for high-order NCS, but additional errors are introduced in the second interpolation.

Interpolation of electron-density values at non-map grid sites is usually required, since the NCS operators will not normally map grid points onto each other. To obtain accurate interpolated values, either a fine grid or a complex interpolation function are required; suitable functions are described in Bricogne (1974) and Cowtan & Main (1998). Solvent flattening and histogram matching are frequently applied after averaging, since histogram matching tends to correct for any smoothing introduced by density interpolation.

In the case of flexible proteins, it may be necessary to average only part of the molecule, in which case the averaging mask will exclude some parts of the unit cell which are indicated as protein by the solvent mask. In other cases, it may be necessary to apply multi-domain averaging; in this case, the protein is divided into rigid domains which can appear in differing orientations. Each domain must then have a separate mask and set of averaging matrices.

Averaging may also be performed across similar molecules in multiple crystal forms (Schuller, 1996); in this case, density modification is performed on each crystal form simultaneously, with averaging of the molecular density across all copies of the molecule in all crystal forms. This is a powerful technique for phase improvement, even when no phasing is available in some crystal forms.

15.1.2.4. Skeletonization

The skeletonization method enhances connectivity in the map. This is achieved by locating ridges of density, constructing a graph of linked peaks, and then building a new map using cylinders of density around the graph peaks.

At worse than atomic resolution, the density peaks for bonded atoms are no longer resolved, and so interpretation of the density in terms of atomic positions involves recognition of common motifs in the pattern of ridges in the density. Skeletonization was a tool developed by Greer (1985) to assist model building by tracing high ridges in the electron density to describe the connectivity in the map.

Skeletonization has more recently been adapted to the problem of density modification (Baker, Bystroff *et al.*, 1993; Bystroff *et al.*, 1993; Wilson & Agard, 1993). A skeleton is constructed by tracing the ridges in the map. The resulting ridges form connected 'trees'. These trees may be pruned to remove small unconnected fragments and break circuits to select for protein-like features. A new map may then be built by building density around the links of the skeleton using the profile of a cylindrically averaged atom at the appropriate resolution.

The skeletonization method has been used to add new features to a partial model of a molecule (Baker, Bystroff *et al.*, 1993). An efficient alternative algorithm for tracing density ridges is given by Swanson (1994).

15.1.2.5. Sayre's equation

Sayre's equation constrains the local shape of electron density. It provides a link between all structure-factor amplitudes and phases. It is an exact equation at atomic resolution in an equal-atom system. It is, therefore, very powerful for phase refinement and extension for small molecules at atomic resolution (Sayre, 1952, 1972, 1974). However, its power diminishes as resolution decreases. It can still be an effective tool for macromolecular phase refinement and extension if the shape function can be modified to accommodate the overlap of atoms at non-atomic resolution (Zhang & Main, 1990*b*).

15.1.2.5.1. Sayre's equation in real and reciprocal space

Sayre's equation (Sayre, 1952, 1972, 1974) expresses the constraint on structure factors when the atoms in a structure are equal and resolved, and the equation has formed the foundation of direct methods. In protein calculations, the resolution is generally too poor for atoms to be resolved, and this is reflected in the bulk of the terms required to calculate the equation for any particular missing structure factor.

For equal and resolved atoms, squaring the electron density changes only the shape of the atomic peaks and not their positions. The original density may therefore be restored by convoluting with some smoothing function, $\psi(\mathbf{x})$, which is a function of atomic shape,

$$\rho(\mathbf{x}) = (V/N)\sum_{\mathbf{y}}\rho^2(\mathbf{y})\psi(\mathbf{x}-\mathbf{y}), \qquad (15.1.2.30)$$

where

$$\psi(\mathbf{x}-\mathbf{y}) = (1/V)\sum_{\mathbf{h}}\theta(\mathbf{h})\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}-\mathbf{y})]. \qquad (15.1.2.31)$$

Here, $\theta(\mathbf{h})$ is the ratio of scattering factors of real, $f(\mathbf{h})$, and 'squared', $g(\mathbf{h})$, atoms, and V is the unit-cell volume, *i.e.*,

$$\theta(\mathbf{h}) = f(\mathbf{h})/g(\mathbf{h}). \qquad (15.1.2.32)$$

Sayre's equation states that the convolution of the squared electron density with a shape function restores the original electron density. It can be seen from equation (15.1.2.30) that Sayre's equation puts constraints on the local shape of electron density. The local shape function is the Fourier transform of the ratio of scattering factors of the real and 'squared' atoms.

Sayre's equation is more frequently expressed in reciprocal space as a system of equations relating structure factors in amplitude and phase:

$$F(\mathbf{h}) = [\theta(\mathbf{h})/V]\sum_{\mathbf{k}}F(\mathbf{k})F(\mathbf{h}-\mathbf{k}). \qquad (15.1.2.33)$$

The reciprocal-space expression of Sayre's equation can be obtained directly from a Fourier transformation of both sides of equation (15.1.2.30) and the application of the convolution theorem.

15.1.2.5.2. The application of Sayre's equation to macromolecules at non-atomic resolution – the θ(h) curve

Sayre's equation is exact for an equal-atom structure at atomic resolution. The reciprocal-space shape function, $\theta(\mathbf{h})$, can be calculated analytically from the ratio of the scattering factors of real and 'squared' atoms, which can both be represented by a Gaussian function. At infinite resolution, we expect $\theta(\mathbf{h})$ to be a spherically symmetric function that decreases smoothly with increased **h**. However, for data at non-atomic resolution, the $\theta(\mathbf{h})$ curve will behave differently because atomic overlap changes the peak shapes. Therefore, a spherical-averaging method is adopted to obtain an estimate of the shape function empirically from the ratio of the observed structure factors and the structure factors from the squared electron density using the formula

$$\theta(s) = V\left\langle F(\mathbf{h})\Big/\sum_{\mathbf{k}}F(\mathbf{k})F(\mathbf{h}-\mathbf{k})\right\rangle_{|\mathbf{h}|}, \qquad (15.1.2.34)$$

where the averaging is carried out over ranges of $|\mathbf{h}|$, *i.e.*, over spherical shells, each covering a narrow resolution range. Here, s represents the modulus of **h**.

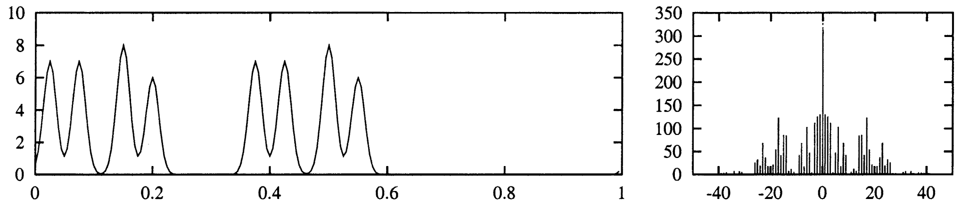
The empirically derived shape function only extends to the resolution of the experimentally observed phases. This is sufficient for phase refinement. However, there are no experimentally observed phases to give the empirical $\theta(s)$ for phase extension. Therefore, a Gaussian function of the form

$$\theta(s) = K\exp(-Bs^2) \qquad (15.1.2.35)$$

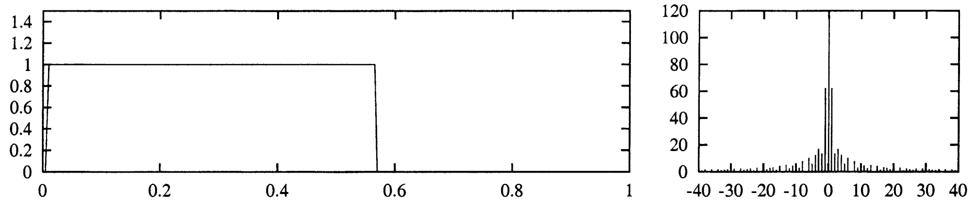
is fitted to the available values of $\theta(s)$, and the parameters K and B are obtained using a least-squares method. The shape function $\theta(s)$ for the resolution beyond that of the observed phases is extrapolated using the fitted Gaussian function. The derivation of the shape function $\theta(s)$ from a combination of spherical averaging and Gaussian extrapolation is the key to the successful application of Sayre's equation for phase improvement at non-atomic resolution (Zhang & Main, 1990*b*).

15.1.2.6. Atomization

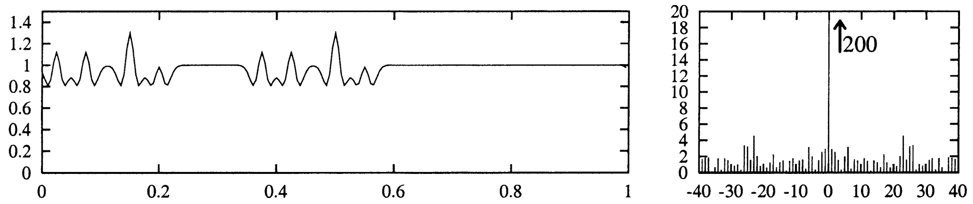
The atomization method uses the fact that the structure underlying the map consists of discrete atoms. It attempts to



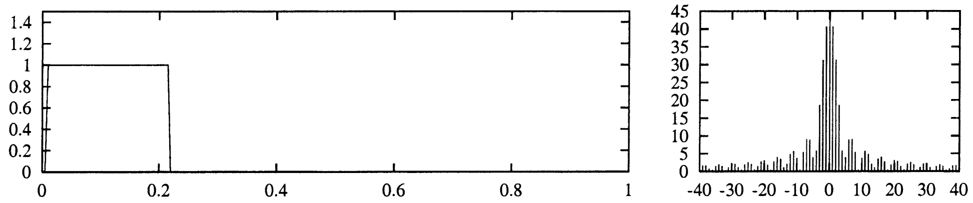
Electron density and structure-factor magnitudes



Solvent-flattening function $g_{\mathrm{sf}}(\mathbf{x})$ and $G_{\mathrm{sf}}(\mathbf{h})$



Histogram-matching function $g_{\mathrm{hm}}(\mathbf{x})$ and $G_{\mathrm{hm}}(\mathbf{h})$



Averaging function $g_{\mathrm{ave}}(\mathbf{x})$ and $G_{\mathrm{ave}}(\mathbf{h})$

Figure 15.1.3.1
The functions $g(\mathbf{x})$ and $G(\mathbf{h})$ for solvent flattening, histogram matching and averaging.

interpret the map by automatically placing atoms and refining their positions.

Agarwal & Isaacs (1977) proposed a method for the extension of phases to higher resolutions by interpreting an electron-density map in terms of 'dummy' atoms. These are so called because at the initial resolution of 3.0 Å, true atom peaks could not be resolved. The placement of 'dummy atoms' is subject to constraints of bonding distance and the number of neighbours. The coordinates and temperature factors of these dummy atoms may then be refined against all the available diffraction amplitudes. Structure factors may then be calculated from the refined coordinates to provide phases for the high-resolution reflections and to improve the phases of the starting set.

The atomization approach has been extended in the *ARP* program (Lamzin & Wilson, 1997) by the use of difference-map criteria to test dummy-atom assignments, with the aim of removing wrong atoms and introducing missing atoms. With modern refinement algorithms, this technique has become very effective for the solution of structures at high resolution from a poor molecular-replacement model, or even directly from an MIR/MAD map.

Map improvement has also been demonstrated at intermediate resolutions by Perrakis *et al.* (1997) using a multi-solution variant of the *ARP* method, and by Vellieux (1998).

The interpretation of an approximately phased map has also been applied very successfully as part of the 'Shake n' Bake' direct-methods procedure (Miller *et al.*, 1993; Weeks *et al.*, 1993). The alternating application of phase refinement by the minimum principle in reciprocal space ('Shake') and atomization in real space ('Bake') has proved to be a very powerful method for solving small protein structures at atomic resolution using only structure-factor amplitudes.

15.1.3. Reciprocal-space interpretation of density modification

Density modification, although mostly performed in real space for ease of application, can be understood in terms of reciprocal-space constraints on structure-factor amplitudes and phases.

Main & Rossmann (1966) showed that the NCS-averaging operation in real space can be expressed in reciprocal space as the convolution of the structure factors and the Fourier transform of the molecular envelope and the NCS matrices. Similarly, the solvent-flattening operation can be considered a multiplication of the map by some mask, $g_{\mathrm{sf}}(\mathbf{x})$, where $g_{\mathrm{sf}}(\mathbf{x}) = 1$ in the protein region and $g_{\mathrm{sf}}(\mathbf{x}) = 0$ in the solvent region. Thus

$$\rho_{\mathrm{mod}}(\mathbf{x}) = g_{\mathrm{sf}}(\mathbf{x}) \times \rho(\mathbf{x}). \qquad (15.1.3.1)$$

This assumes that the solvent level is zero, which can be achieved by suitable adjustment of the $F(000)$ term.

If we transform this equation to reciprocal space, then the product becomes a convolution; thus

$$F_{\mathrm{mod}}(\mathbf{h}) = (1/V)\sum_{\mathbf{k}} G_{\mathrm{sf}}(\mathbf{k})F(\mathbf{h} - \mathbf{k}), \qquad (15.1.3.2)$$

where $G_{\rm sf}(\mathbf{k})$ is the Fourier transform of the mask $g_{\rm sf}(\mathbf{x})$. The solvent mask $g_{\rm sf}(\mathbf{x})$ shows the outline of the molecule with no internal detail, so must be a low-resolution image. Therefore, all but the lowest-resolution terms of $G_{\rm sf}$ will be negligible.

The convolution expresses the relationship between phases in reciprocal space from the constraint of solvent flatness in real space. Since only the terms near the origin of $G_{\rm sf}$ are nonzero, the convolution can only relate phases that are local to each other in reciprocal space. Thus, it can only provide phase information for structure factors near the current phasing resolution limit.

This reasoning may also be applied to other density modifications. Histogram matching applies a nonlinear rescaling to the current density in the protein region. The equivalent multiplier, $g_{\rm hm}(\mathbf{x})$, shows variations of about 1.0 that are related to the features in the initial map. The function $G_{\rm hm}(\mathbf{h})$ for histogram matching is, therefore, dominated by its origin term, but shows significant features to the same resolution as the current map or further, as the density rescaling becomes more nonlinear. Histogram matching can therefore give phase indications to twice the resolution of the initial map or beyond, although phase indications will be weak and contain errors related to the level of error in the initial map.

$$\rho_{\rm mod}(\mathbf{x}) = g_{\rm ncs}(\mathbf{x})(1/N_{\rm ncs})\sum_i \rho_i(\mathbf{x}). \tag{15.1.3.3}$$

Averaging may be described as the summation of a number of reoriented copies of the electron density within the region of the averaging mask (Main & Rossmann, 1966), *i.e.* where $\rho_i(\mathbf{x})$ is the initial density, $\rho(\mathbf{x})$, transformed by the ith NCS operator and $g_{\rm ncs}(\mathbf{x})$ is the mask of the molecule to be averaged. This summation is repeated for each copy of the molecule in the whole unit cell. The reciprocal-space averaging function, $G_{\rm ncs}(\mathbf{h})$, is the Fourier transform of a mask, as for solvent flattening, but since the mask covers only a single molecule, rather than the molecular density in the whole unit cell, the extent of $G_{\rm ncs}(\mathbf{h})$ in reciprocal space is greater.

Sayre's equation is already expressed as a convolution, although in this case the function $G(\mathbf{h})$ is given by the structure factors $F(\mathbf{h})$ themselves. It is, therefore, the most powerful method for phase extension. However, as resolution decreases, more of the reflections required to form the convolution are missing, and the error increases.

The functions $g(\mathbf{x})$ and $G(\mathbf{h})$ for these density modifications are illustrated in Fig. 15.1.3.1 for a simple one-dimensional structure.

15.1.4. Phase combination

Phase combination is used to filter the noise in the modified phases and eliminate the incorrect component of the modified phases through a statistical process. The observed structure-factor amplitudes are used to estimate the reliability of the phases after density modification. The estimated probability of the modified phases is combined with the probability of observed phases to produce a more reliable phase estimate,

$$P_{\rm new}[\varphi(\mathbf{h})] = P_{\rm obs}[\varphi(\mathbf{h})]P_{\rm mod}[\varphi(\mathbf{h})]. \tag{15.1.4.1}$$

Once a modified map has been obtained, modified phases and amplitudes may be derived from an inverse Fourier transform. The modified phases are normally combined with the initial phases by multiplication of their probability distributions. The probability distribution for the experimentally observed phases is usually described in terms of a best phase and figure of merit (Blow & Rossmann, 1961) or by Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970). In order to estimate a unimodal probability distribution for the modified phase, some estimate of the associated error must be made; this is usually achieved using the Sim weighting scheme (Sim, 1959).

Recombination with the initial phases assumes independence between the initial and modified phases and is a source of difficulties. However, in the absence of some form of phase constraint, most density-modification constraints are too weak to guarantee convergence to a reasonable solution. The exception is when high-order NCS is present; in this case, the combination of NCS and observed amplitudes is sufficient to determine the phases (Chapman *et al.*, 1992; Tsao *et al.*, 1992), and phase combination may be omitted; however, weighting of the phases is still necessary. In this case, it is also possible to restore missing reflections in both amplitude and phase.

15.1.4.1. Sim and σ_a weighting

The phase probability distribution for the density-modified phase is conventionally generated under assumptions that were made for the combination of a partial atomic model with experimental data. It assumes that the calculated amplitudes and phases arise from a density map in which some atoms are present and correctly positioned, and the remainder are completely absent (Sim, 1959). Thus, the difference between the true structure factor and the calculated value must be the effective structure factor due to the missing density alone. If the phase of this quantity is random and the amplitude is drawn from a Wilson distribution (Wilson, 1949), the following expression is obtained:

$$P_{\rm mod}(\varphi) = \exp[A\cos\varphi + B\sin\varphi], \tag{15.1.4.2}$$

where

$$\begin{aligned} A &= X\cos\varphi_{\rm exp} \\ B &= X\sin\varphi_{\rm exp} \end{aligned} \tag{15.1.4.3}$$

and

$$X = 2|F_{\rm exp}||F_{\rm mod}|/\Sigma_Q, \tag{15.1.4.4}$$

where Σ_Q is the variance parameter in the Wilson distribution for the missing part of the structure. The figure of merit, w, can be derived from

$$w = I_1(X)/I_0(X), \tag{15.1.4.5}$$

where I_0 and I_1 are zero- and first-order modified Bessel functions. A similar argument follows for centric reflections.

The error estimate for the phase depends on the effective amount of missing structure that is estimated on the basis of the agreement of the modified amplitudes with their measured values, where Σ_Q may be estimated by a number of means, for example (Bricogne, 1976),

$$\Sigma_Q = \langle |F_{\rm obs}|^2 - |F_{\rm mod}|^2 \rangle, \tag{15.1.4.6}$$

where the average is normally taken over all reflections at a particular resolution. A more sophisticated approach is the σ_a method of Read (1986), which allows for errors in the atomic model and has also been used in density modification (Chapter 15.2).

Although these approaches have been applied with some success, the assumption in equation (15.1.4.1) that the density-modified amplitudes and phases are independent of the initial values is invalid. Since the density constraints are typically underdetermined, it is possible to achieve an arbitrarily good agreement between the model amplitudes and their observed values without improving the phases. As a result, phase weights from density modification are typically overestimated.

This problem has traditionally been addressed by limiting the number of cycles of density modification in which weakly phased reflections are included. Typically, density modification is started with only some subset of the data, such as those reflections well phased from MIR data. Only these reflections are included in the phase recombination, with other reflections set to zero. As the calculation progresses, more reflections are introduced until all the data are included. The figures of merit of reflections that undergo fewer cycles of phase recombination will be correspondingly smaller (*e.g.* Leslie, 1987; Zhang & Main, 1990*a*). In averaging calculations where considerable phase information is available from high-order NCS, it is still typically necessary to perform phase extension over hundreds of cycles and to add a very thin resolution shell of new reflections at each cycle.

The phases and figure of merit generated from density modification are more suited to the calculation of weighted F_o maps than $2mF_o - F_c$ maps. The $2mF_o - F_c$ map is designed to aid the structure completion from a partial model (Main, 1979). The $2mF_o - F_c$ map will restore features missing from the current model at full weight if the following conditions are fulfilled. First, the model phases must be close to their true values. Secondly, the difference between the model and observed amplitudes is a good indicator of the phase error and the difference between the calculated and observed amplitudes decreases as the phases approach their true values. Neither of these assumptions are necessarily true for density modification, since it may be applied to very poor maps with almost random phases, and under most density-modification schemes the structure-factor amplitudes may be over-fitted to the observed values.

15.1.4.2. *Reflection omit*

The modified map may be made more independent of the original map, as was assumed when multiplying the phase probability distributions in equation (15.1.4.1), through a reciprocal-space analogue of the omit map, the reflection-omit method.

The reflections are divided into (typically 10 or 20) sets and density-modification calculations are performed, excluding each set in turn from the calculation of the starting map, in a manner similar to a free-*R*-value calculation (Brünger, 1992). Density modification is applied to each map in turn, and the modified reflections from each of the free sets are combined to give a new, complete data set. This data set should be less dependent on the original amplitudes; therefore, the amplitudes may be expected to give a better indication of the quality of the modified phases.

The resulting maps obtained using solvent flattening and/or histogram matching are dramatically improved using the reflection-omit method (Cowtan & Main, 1996). In the case of averaging calculations, however, the reflection-omit approach makes little difference, since omitted reflections tend to be restored through noncrystallographic symmetry relationships to other regions of reciprocal space. It is possible that further improvements may be achieved by selecting reflection sets that approximately obey the NCS relationships.

15.1.4.3. *The γ correction and solvent flipping*

Abrahams & Leslie (1996) have shown that solvent flipping is dramatically more effective as a density modification than solvent flattening. This may be shown to be theoretically equivalent to performing a reflection-omit calculation for each reflection individually (Abrahams, 1997).

Solvent flattening is represented in reciprocal space by convolution of the structure factors with a function, $G(\mathbf{h})$, as shown in equation (15.1.3.2). If the origin term of G is set to zero, then the modified structure factor, $F_{\rm mod}(\mathbf{h})$, will depend on the values of all the structure factors except itself; this is equivalent to performing a reflection-omit calculation with that reflection alone omitted.

Let the origin-removed G be called $G_\gamma(\mathbf{h})$ and its Fourier transform $g_\gamma(\mathbf{x})$:

$$G_\gamma(\mathbf{h}) = \begin{cases} 0, & \mathbf{h} = 0 \\ G(\mathbf{h}), & \mathbf{h} \neq 0 \end{cases}, \qquad (15.1.4.7)$$

then

$$g_\gamma(\mathbf{x}) = g(\mathbf{x}) - \overline{g(\mathbf{x})}. \qquad (15.1.4.8)$$

The convolution of the reflection data with $G_\gamma(\mathbf{h})$ is equivalent to performing a reflection-omit calculation, omitting every reflection in turn. However, the convolution may still be performed in real space; thus, the full omit calculation becomes a simple multiplication of the map by $g_\gamma(\mathbf{x})$:

$$\rho_{\rm mod}(\mathbf{x}) = g_\gamma(\mathbf{x}) \times \rho(\mathbf{x}). \qquad (15.1.4.9)$$

In a solvent-flattening calculation, $g_\gamma(\mathbf{x})$ will be equal to $g(\mathbf{x})$ minus the fraction of the cell that is protein. In the case of a cell with 50% solvent, $g_\gamma(\mathbf{x})$ has a value of 0.5 in the protein and -0.5 in the solvent. Multiplication of the map by this function results in flipping of the solvent.

If the origin term of the G function, γ, can be determined, then the flipping calculation may alternatively be performed by subtracting a copy of the initial map scaled by γ from the modified map. This is the γ correction of Abrahams (1997). This approach may be generalized to arbitrary density-modification methods by use of the perturbation γ (Cowtan, 1999). In this approach, a random perturbation is applied to the starting data. Density modification is applied to both the perturbed and unperturbed maps. The relative size of the perturbation signal in the modified map gives an estimate for γ. The perturbation γ provides effective bias correction for any combination of solvent flattening, histogram matching and averaging. γ may also be estimated as a function of resolution, allowing successful application to multi-resolution modification and possibly atomization as well.

15.1.4.4. *Resolution extrapolation*

The Fourier coefficients of the density-modified map include nonzero contributions for reflections that were not present in the original electron-density map. These values are commonly used to restore the values of reflections that were missing from the original data (including low-resolution reflections falling behind the backstop). However, they have recently also been employed to extrapolate structure factors beyond the resolution of the observed data (Caliandro *et al.*, 2005; Usón *et al.*, 2007).

In the scheme of Usón *et al.*, the Fourier coefficients of the density-modified map are normalized to obey a Wilson distribution (calculated using the overall temperature factor of the observed data). These are then used in the calculation of the next electron-density map. At high resolution (1.35 Å in the case reported), this procedure leads to both improved phases for the observed structure factors at lower resolutions and good estimates for the extrapolated structure factors – these can be judged by comparison with the calculated values obtained from the final refined model. The resulting electron-density map obtained using density modification alone can rival the quality of the final 'best' map obtained from refinement.

The theoretical basis for the effectiveness of this approach has yet to be fully explained. Current implementations are mainly effective at high resolution, where the extrapolated resolution can approach atomic resolution.

15.1.5. Combining constraints for phase improvement

The chemical and physical information of the underlying structure that the electron density represents serves as constraints on the phases. For small molecules, the constraints of positivity and atomicity are sufficient to solve the phase problem *ab initio* (Hauptman, 1986; Karle, 1986; Woolfson, 1987), because crystals of small molecules generally diffract to atomic resolution. However, no single constraint at our disposal is powerful enough to render the macromolecular phase problem determinable, because macromolecule crystals rarely diffract to atomic resolution. Therefore, individual constraints are combined to produce a more powerful density-modification protocol. This is because these constraints represent different characteristic features of the electron density and they contain independent phasing information.

The phasing power of a method increases with the number of independent constraints employed, the number of density points affected and the amplitude of changes imposed on the electron density. It also depends on the physical nature and accuracy of the constraints and how the constraints are applied. One obvious way of implementing several constraints is to apply them one after the other to the electron density. This sequential application, although easy to implement, suffers some drawbacks. The cyclic application of all constraints may not converge easily, since some constraints may contain contradicting information as to how the density should be modified. An alternative way of implementing various constraints is simultaneous application. The density solution that satisfies all the constraints is obtained by a global minimization procedure (Main, 1990*b*; Zhang & Main, 1990*b*).

15.1.5.1. The system of nonlinear constraint equations

The constraints used in *SQUASH/DM* can be divided into three categories. The first category comprises the linear constraints, such as solvent flatness, density histogram and equal molecules. The second category comprises the nonlinear constraints, such as the local shape of electron density as expressed in Sayre's equation. The third category comprises the available structural data, such as the observed structure-factor amplitudes and the experimental phases. The first and second categories of constraints are used to solve new electron-density values. The third category of constraints is used as a means to filter the modified phases.

The modification to the density value at a grid point by a linear constraint is independent of the values at other grid points. These constraints include solvent flattening, histogram matching and molecular averaging. These density-modification methods construct an improved map directly from an initial density map as expressed by

$$\rho(\mathbf{x}) = H(\mathbf{x}), \qquad (15.1.5.1)$$

where $H(\mathbf{x})$ is the target electron density produced by these linear constraints.

The new electron density that satisfies both the linear constraints represented by equation (15.1.5.1) and the nonlinear constraints expressed by Sayre's equation (15.1.2.30) can be obtained by solving the systems of simultaneous equations (Zhang & Main, 1990*b*)

$$\begin{cases} (V/N)\sum_{\mathbf{y}}\rho^2(\mathbf{y})\psi(\mathbf{x}-\mathbf{y}) - \rho(\mathbf{x}) = 0 \\ H(\mathbf{x}) - \rho(\mathbf{x}) = 0 \end{cases}. \qquad (15.1.5.2)$$

Equation (15.1.5.2) represents a system of nonlinear simultaneous equations with as many unknowns as the number of grid points in the asymmetric unit of the map and with twice as many equations as unknowns. The functions $H(\mathbf{x})$ and $\psi(\mathbf{x}-\mathbf{y})$ are both known. The least-squares solution, using either the full matrix or the diagonal approximation, is obtained using the Newton–Raphson technique with fast Fourier transforms, as described in the next section (Main, 1990*b*).

15.1.5.2. Least-squares solution to the system of nonlinear constraint equations

For a system of nonlinear equations of electron density,

$$\mathbf{F}(\rho(\mathbf{x})) = \mathbf{0}, \qquad (15.1.5.3)$$

where

$$\mathbf{F}(\rho(\mathbf{x})) = \left[F_1(\rho(\mathbf{x}))\ F_2(\rho(\mathbf{x}))\ \ldots\ F_m(\rho(\mathbf{x}))\right]^T,$$
$$\rho(\mathbf{x}) = \left[\rho_1\ \rho_2\ \ldots\ \rho_n\right]^T,$$

$\mathbf{0}$ is a null vector, n is the number of grid points and m is the number of equations, the Newton–Raphson method of solution is to find a set of shifts, $\delta\rho(\mathbf{x})$ to $\rho(\mathbf{x})$, through a system of linear equations,

$$\mathbf{J}\delta\rho(\mathbf{x}) = -\varepsilon, \qquad (15.1.5.4)$$

where $\mathbf{J}$ is a matrix of partial derivatives of $\mathbf{F}$ with respect to $\rho(\mathbf{x})$ and is called the Jacobian matrix,

$$\mathbf{J} = \begin{bmatrix} \frac{\partial F_1}{\partial \rho_1} & \frac{\partial F_1}{\partial \rho_2} & \cdots & \frac{\partial F_1}{\partial \rho_n} \\ \frac{\partial F_2}{\partial \rho_1} & \frac{\partial F_2}{\partial \rho_2} & \cdots & \frac{\partial F_2}{\partial \rho_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial \rho_1} & \frac{\partial F_m}{\partial \rho_2} & \cdots & \frac{\partial F_m}{\partial \rho_n} \end{bmatrix}, \qquad (15.1.5.5)$$

ε is a vector of residuals to equation (15.1.5.3) for a trial solution, $\rho(\mathbf{x})$, and $\delta\rho(\mathbf{x})$ is a vector of shifts to the density. Hence, the solution for $\rho(\mathbf{x})$ is achieved in an iterative manner,

$$\rho^{i+1}(\mathbf{x}) = \rho^i(\mathbf{x}) + \delta\rho(\mathbf{x}). \qquad (15.1.5.6)$$

Therefore, the problem of solving a system of nonlinear equations (15.1.5.3) is transformed into solving a system of linear equations (15.1.5.4), which forms one cycle of Newton–Raphson iteration.

If there are more equations than unknowns ($m > n$), the unknowns are obtained through a least-squares solution to equations (15.1.5.4),

$$\mathbf{J}^T\mathbf{J}\delta\rho(\mathbf{x}) = -\mathbf{J}^T\varepsilon. \qquad (15.1.5.7)$$

Theoretically, the above system of equations could be solved by matrix multiplication and inversion, *i.e.*

$$\delta\rho(\mathbf{x}) = -\left(\mathbf{J}^T\mathbf{J}\right)^{-1}\mathbf{J}^T\varepsilon. \qquad (15.1.5.8)$$

However, the amount of calculation involved in setting up the normal matrix of least squares is huge for the problem presented by protein structures. This can be completely avoided by using the conjugate-gradient technique for solving the system of linear equations.

15.1.5.2.1. The conjugate-gradient method

The conjugate-gradient method does not require the inversion of the normal matrix, and therefore the solution to a large system of linear equations can be achieved very quickly.

Starting from a trial solution to equations (15.1.5.4), such as a null vector,

$$\delta\rho_0(\mathbf{x}) = \mathbf{0}, \qquad (15.1.5.9)$$

the initial residual is

$$\mathbf{r}_0 = -\mathbf{J}^T(\varepsilon - \mathbf{J}\delta\rho_0(\mathbf{x})) \qquad (15.1.5.10)$$

and the initial search step is

$$\mathbf{p}_0 = \mathbf{r}_0. \qquad (15.1.5.11)$$

The iterative process is as follows. The new shift to the density is

$$\delta\rho_{k+1}(\mathbf{x}) = \delta\rho_k(\mathbf{x}) + \alpha_k\mathbf{p}_k, \qquad (15.1.5.12)$$

where

$$\alpha_k = \mathbf{r}_k^T\mathbf{p}_k/\mathbf{q}_k^T\mathbf{q}_k \qquad (15.1.5.13)$$

and

$$\mathbf{q}_k = \mathbf{J}\mathbf{p}_k. \qquad (15.1.5.14)$$

The new residual is

$$\mathbf{r}_{k+1} = \mathbf{r}_k - \alpha_k\mathbf{s}_k, \qquad (15.1.5.15)$$

where

$$\mathbf{s}_k = \mathbf{J}^T\mathbf{q}_k. \qquad (15.1.5.16)$$

The next search step which conjugates with the residual is

$$\mathbf{p}_{k+1} = \mathbf{r}_{k+1} + \beta_k\mathbf{p}_k, \qquad (15.1.5.17)$$

where

$$\beta_k = -\mathbf{r}_{k+1}^T\mathbf{s}_k/\mathbf{q}_k^T\mathbf{q}_k. \qquad (15.1.5.18)$$

The process is iterated by increasing k until convergence is reached, when

$$\left|\mathbf{r}_{k+1} - \mathbf{r}_k\right| \Rightarrow 0.$$

The number of iterations required for an exact solution is equal to the number of unknowns, because the search vector at each step is orthogonal with all the previous steps. However, a very satisfactory solution can normally be reached after very few iterations. This makes the conjugate-gradient method a very efficient and fast procedure for solving a system of equations. Note that the normal matrix never appears explicitly, although it is implicit in (15.1.5.10) and (15.1.5.16). The inversion of the normal matrix and matrix multiplication is completely avoided. Most of the calculation comes from the formation of the matrix-vector products in (15.1.5.10), (15.1.5.14) and (15.1.5.16). These can be expressed as convolutions and can be performed using FFTs, thus saving considerably more time.

The solution to $\delta\rho(\mathbf{x})$ at the end of conjugate-gradient iteration is substituted into equation (15.1.5.6) to get a new solution for $\rho(\mathbf{x})$. The solution to the system of nonlinear equations (15.1.5.3) is obtained when the Newton–Raphson iteration has reached convergence.

15.1.5.2.2. The full-matrix solution

The equations to be solved for the electron-density shifts, $\delta\rho(\mathbf{x})$, are from the Jacobian of equation (15.1.5.2),

$$\left\{\begin{array}{l}(2V/N)\sum\limits_{\mathbf{y}}\rho(\mathbf{y})\psi(\mathbf{x}-\mathbf{y}) - \delta\rho(\mathbf{x}) = \Delta\rho(\mathbf{x})\\ \delta\rho(\mathbf{x}) = \Delta H(\mathbf{x})\end{array}\right., \qquad (15.1.5.19)$$

where $\Delta\rho(\mathbf{x})$ is the residual to Sayre's equation,

$$\Delta\rho(\mathbf{x}) = \rho(\mathbf{x}) - (V/N)\sum_{\mathbf{y}}\rho^2(\mathbf{y})\psi(\mathbf{x}-\mathbf{y}), \qquad (15.1.5.20)$$

and $\Delta H(\mathbf{x})$ is the residual to the linear density-modification equations,

$$\Delta H(\mathbf{x}) = H(\mathbf{x}) - \rho(\mathbf{x}). \qquad (15.1.5.21)$$

Starting from a trial solution of $\delta\rho_0(\mathbf{x}) = \mathbf{0}$, the initial residual vector is

$$\begin{aligned}\mathbf{r}_0(\mathbf{x}) = {} & (2/V)\rho(\mathbf{x})\sum_{\mathbf{h}}\theta(\bar{\mathbf{h}})\Delta F(\mathbf{h})\exp(-2\pi i\mathbf{hx})\\ & - \Delta\rho(\mathbf{x}) + \Delta H(\mathbf{x}),\end{aligned} \qquad (15.1.5.22)$$

where

$$\Delta F(\mathbf{h}) = F(\mathbf{h}) - \theta(\mathbf{h})G(\mathbf{h}), \qquad (15.1.5.23)$$

$$G(\mathbf{h}) = (V/N)\sum_{\mathbf{y}}\rho^2(\mathbf{y})\exp(2\pi i\mathbf{hy}) \qquad (15.1.5.24)$$

and

$$\Delta\rho(\mathbf{x}) = (1/V)\sum_{\mathbf{h}}\Delta F(\mathbf{h})\exp(-2\pi i\mathbf{hx}). \qquad (15.1.5.25)$$

Thus, only three FFTs are required to calculate the initial residual. The residual of Sayre's equation is given in equation (15.1.5.23).

The calculation of $\mathbf{q}_k$ in equation (15.1.5.14) is achieved in a similar manner using FFTs,

$$\begin{aligned}\mathbf{q}_k = \mathbf{J}\mathbf{p}_k &= \left\{\begin{matrix}(1/V)\sum_{\mathbf{h}}[2a(\mathbf{h})\theta(\mathbf{h}) - b(\mathbf{h})]\exp(-2\pi i\mathbf{hx})\\ \hline p_k(\mathbf{x})\end{matrix}\right\}\\ &= \left[\begin{matrix}Q_k(\mathbf{x})\\ \hline p_k(\mathbf{x})\end{matrix}\right],\end{aligned} \qquad (15.1.5.26)$$

where the vector is partitioned as shown above, and

$$a(\mathbf{h}) = (V/N)\sum_{\mathbf{y}}\rho(\mathbf{y})p_k(\mathbf{y})\exp(2\pi i\mathbf{hy}), \qquad (15.1.5.27)$$

$$b(\mathbf{h}) = (V/N)\sum_{\mathbf{y}} p_k(\mathbf{y})\exp(2\pi i\mathbf{hy}). \qquad (15.1.5.28)$$

Similarly, vector $\mathbf{s}_k$ in equation (15.1.5.16) is obtained from

$$\begin{aligned}\mathbf{s}_k = \mathbf{J}^T\mathbf{q}_k = {} & (2/V)\rho(\mathbf{x})\sum_{\mathbf{h}}\theta(\bar{\mathbf{h}})[2a(\mathbf{h})\theta(\mathbf{h}) - b(\mathbf{h})]\exp(-2\pi i\mathbf{hx})\\ & -Q_k(\mathbf{x}) + p_k(\mathbf{x}),\end{aligned} \qquad (15.1.5.29)$$

where $Q_k(\mathbf{x})$ is defined in equation (15.1.5.26).

The remaining calculations in equations (15.1.5.12), (15.1.5.13), (15.1.5.15), (15.1.5.17) and (15.1.5.18) require either the inner product of a pair of vectors or a linear combination of vectors, both of which are very quick to calculate. Each iteration of the conjugate gradient requires four FFTs, as described in equations (15.1.5.26)–(15.1.5.29).

15.1.5.2.3. The diagonal approximation

The full-matrix solution to equation (15.1.5.4) requires a significant amount of computing, although it can be achieved using FFTs. The diagonal approximation to the normal matrix has been used as an alternative method of solution to the electron-density shift in equation (15.1.5.4) (Main, 1990*b*). As with the full-matrix calculation, it can be done entirely by FFTs and a linear combination of vectors.

The diagonal element of the normal matrix, $\mathbf{J}^T\mathbf{J}$, in equation (15.1.5.7) is

$$d_0(\mathbf{x}) = (4/N)\rho(\mathbf{x})\left[\rho(\mathbf{x})\sum_{\mathbf{h}}|\theta(\mathbf{h})|^2 - \sum_{\mathbf{h}}\theta(\mathbf{h})\right] + 2. \qquad (15.1.5.30)$$

The right-hand side of equation (15.1.5.7), $-\mathbf{J}^T\varepsilon(\mathbf{x})$, is identical to the residual vector, $r_0(\mathbf{x})$, which can be calculated from equation (15.1.5.22). Therefore, the solution to the electron-density shift, $\delta\rho(\mathbf{x})$, can be calculated from

$$\delta\rho(\mathbf{x}) = r_0(\mathbf{x})/d_0(\mathbf{x}). \qquad (15.1.5.31)$$

Compared with the full-matrix solution, all the calculations involved in between equations (15.1.5.12) and (15.1.5.18) and the subsequent iterations are spared in the diagonal approximation. This makes calculation by the diagonal approximation much faster than by the full-matrix method.

15.1.6. Statistical density-modification methods

Statistical density-modification methods arise from a reinterpretation of the problem of phase improvement in statistical terms, and as a result reduce the problems of bias associated with the classical density-modification methods described above.

This is achieved by expressing all information about probable electron-density values in the map in terms of probability distributions, with a probability distribution of electron-density values assigned to each point in the electron-density map. The probable electron-density values at any point in the map may depend, for example, on whether that point lies in the solvent or protein region, and on any NCS relationships with other regions of the map. These distributions are then used to infer corresponding phase probability distributions in reciprocal space, which may then be combined with any existing phase information. This avoids working with a single map representing a single sample from the phase probability distributions.

This formulation in turn weakens the link between the newly introduced information and the initial phase probability distributions, thus reducing the bias introduced in a single cycle of phase improvement. Since the current 'best' map is not used as a starting map to be modified, the phase probability distributions from which the 'best' map is derived are not directly included in any new phase information. The only way in which the current phases are used is in the classification of the asymmetric unit into regions of different density types, *e.g.* solvent and protein.

The result of these changes is that statistical density-modification techniques lead to reduced phase bias and more realistic estimates of the figures of merit. The resulting method has been implemented in the *RESOLVE* software (Terwilliger, 1999). In addition to its application to conventional density-modification problems, it has been particularly effective in removing bias from maps phased from an atomic model through the 'prime-and-switch' approach (Terwilliger, 2004). The statistical approach to density modification requires substantially more computation that the simpler classical methods.

15.1.7. Example

To demonstrate the effect of different constraints on phase improvement, various density-modification techniques were applied to an MIR data set for which the refined structure coordinates are available. The test structure is 5-carboxymethyl-2-hydroxymuconate isomerase, solved by Wigley *et al.* (1989). MIR phases were available to 3.7 Å, with SIR information to 2.6 Å. Density modification was used to improve and extend phases to the limit of the data at 2.1 Å. The structure includes threefold noncrystallographic symmetry.

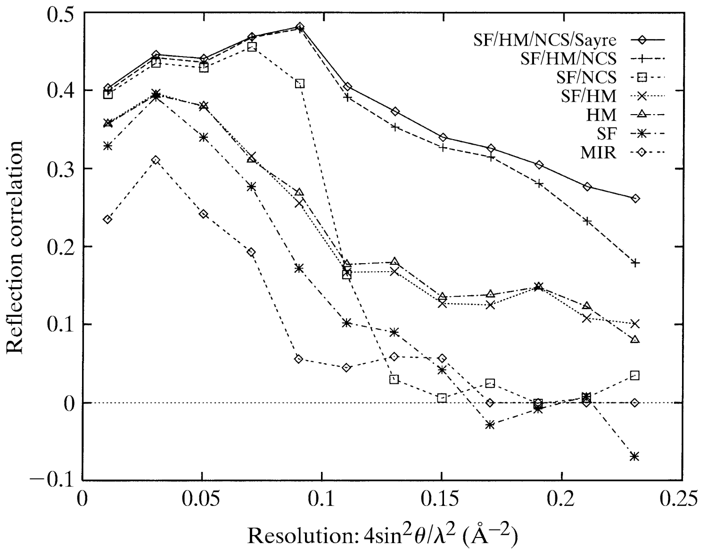


Figure 15.1.7.1

Phase correlations after different combinations of density modifications.

The MIR and density-modified phases are compared by plotting the mean of the cosine of the phase error, weighted by the figure of merit and structure-factor amplitude, as a function of resolution (Zhang *et al.*, 1997),

$$C_f = \langle w|F|^2\cos(\varphi - \varphi_0)\rangle / \left(\langle w^2|F|^2\rangle\langle|F|^2\rangle\right)^{1/2}. \qquad (15.1.7.1)$$

This phase correlation over all reflections is equivalent to map correlation. The results of density modification by various techniques, using the reflection-omit method for phase combination, are shown in Fig. 15.1.7.1.

Solvent flattening alone has slightly improved the phases at low resolution but has not lead to significant phase extension. The solvent-flattening function in Fig. 15.1.3.1 only has nonzero amplitudes close to the origin. It relates structure factors only in a very thin resolution shell. Therefore, solvent flattening is weak on phase extension.

Histogram matching alone improves the low-resolution phases and gives significant phase extension to higher resolutions. The histogram-matching function in Fig. 15.1.3.1 showed much stronger high-resolution amplitudes. Therefore, it could relate structure factors in a larger resolution shell. Moreover, there is always an ideal histogram specified at a given target resolution for phase extension. These two reasons combined make histogram matching a more powerful technique in phase extension than solvent flattening.

The combination of histogram matching and solvent flattening is slightly more powerful than histogram matching alone; since histogram matching sharpens the protein density, it implies an element of solvent flattening. Solvent flattening and averaging give a significant improvement at low resolution, but little phase extension. Averaging is powerful for phase refinement, but is weak for phase extension if no special precautions are taken. If there are flexible loop regions on the protein surface, these regions should be excluded from the molecular mask for averaging. The phasing power of averaging weakens at high resolution when the differences between NCS-related molecules become significant. Solvent flattening, histogram matching and averaging combined give a dramatic improvement at all resolutions. The addition of Sayre's equation gives a slight further improvement at high resolution.

Sayre's equation is very effective for phase refinement and extension at atomic or near atomic resolution. It becomes ineffective at low resolution or when the initial map is poor. Under these circumstances, it is better to apply other density-modification methods first to refine the phases and extend them to a higher resolution before Sayre's equation is applied. Sayre's equation also decreases in power as the solvent content increases, since it is only applicable to the protein regions of the map.

The fact that the best results were obtained when all the constraints were combined indicates that each constraint contains some degree of independent phasing information. Moreover, it also suggests that the strengths of these constraints are complementary. Each constraint, when applied in isolation, may introduce systematic errors that are difficult to overcome when a different constraint is subsequently applied. This problem is greatly reduced when the constraints are applied simultaneously and the combined process iterates much further towards the desired density map.

Density-modification methods have become sufficiently powerful that it is possible to solve structures from comparatively poor initial maps. This has reduced the amount of effort required to find more heavy-atom derivatives and to collect additional diffraction data sets. Density modification may simplify the process of map interpretation, even when good phase information is available. Density modification can also be used to obtain phases *ab initio* when high-order noncrystallographic symmetry is present.

KYJZ acknowledges the US National Institutes of Health for support (grant GM55663). KDC acknowledges the UK BBSRC for support (grant 87/B03785). Some of the material used in this article is reprinted from Cowtan & Zhang (1999) with permission from Elsevier Science.

References

Abrahams, J. P. (1997). *Bias reduction in phase refinement by modified interference functions: introducing the γ correction. Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F_1 ATPase. Acta Cryst.* D**52**, 30–42.

Agarwal, R. C. & Isaacs, N. W. (1977). *Method for obtaining a high resolution protein map starting from a low resolution map. Proc. Natl Acad. Sci. USA*, **74**, 2835–2839.

Baker, D., Bystroff, C., Fletterick, R. J. & Agard, D. A. (1993). *PRISM: topologically constrained phase refinement for macromolecular crystallography. Acta Cryst.* D**49**, 429–439.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Uniqueness and the ab initio phase problem in macromolecular crystallography. Acta Cryst.* D**49**, 186–192.

Bhat, T. N. & Blow, D. M. (1982). *A density-modification method for the improvement of poorly resolved protein electron-density maps. Acta Cryst.* A**38**, 21–29.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method. Acta Cryst.* **14**, 1195–1202.

Bricogne, G. (1974). *Geometric sources of redundancy in intensity data and their use for phase determination. Acta Cryst.* A**30**, 395–405.

Bricogne, G. (1976). *Methods and programs for direct-space exploitation of geometric redundancies. Acta Cryst.* A**32**, 832–847.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Bystroff, C., Baker, D., Fletterick, R. J. & Agard, D. A. (1993). *PRISM: application to the solution of two protein structures. Acta Cryst.* D**49**, 440–448.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005). *Phasing at resolution higher than the experimental resolution. Acta Cryst.* D**61**, 556–565.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Ab initio phase determination for spherical viruses: parameter determination for spherical-shell models. Acta Cryst.* A**48**, 301–312.

Cowtan, K. D. (1999). *Error estimation and bias correction in phase-improvement calculations. Acta Cryst.* D**55**, 1555–1567.

Cowtan, K. D. & Main, P. (1993). *Improvement of macromolecular electron-density maps by the simultaneous application of real and reciprocal space constraints. Acta Cryst.* D**49**, 148–157.

Cowtan, K. D. & Main, P. (1996). *Phase combination and cross validation in iterated density-modification calculations. Acta Cryst.* D**52**, 43–48.

Cowtan, K. D. & Main, P. (1998). *Miscellaneous algorithms for density modification. Acta Cryst.* D**54**, 487–493.

Cowtan, K. D. & Zhang, K. Y. J. (1999). *Density modification for macromolecular phase improvement. Prog. Biophys. Mol. Biol.* **72**, 245–270.

Crowther, R. A. & Blow, D. M. (1967). *A method of positioning a known molecule in an unknown crystal structure. Acta Cryst.* **23**, 544–548.

Greer, J. (1985). *Computer skeletonization and automatic electron-density map analysis.* In *Diffraction Methods for Biological Macromolecules*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, Vol. 115, pp. 206–224. Orlando: Academic Press.

Harrison, R. W. (1988). *Histogram specification as a method of density modification. J. Appl. Cryst.* **21**, 949–952.

Hauptman, H. (1986). *The direct methods of X-ray crystallography. Science*, **233**, 178–183.

Hendrickson, W. A., Klippenstein, G. L. & Ward, K. B. (1975). *Tertiary structure of myohemerythrin at low resolution. Proc. Natl Acad. Sci. USA*, **72**, 2160–2164.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Hoppe, W. & Gassmann, J. (1968). *Phase correction, a new method to solve partially known structures. Acta Cryst.* B**24**, 97–107.

Karle, J. (1986). *Recovering phase information from intensity data. Science*, **232**, 837–843.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography. Methods Enzymol.* **277**, 269–305.

Leslie, A. G. W. (1987). *A reciprocal-space method for calculating a molecular envelope using the algorithm of B. C. Wang. Acta Cryst.* A**43**, 134–136.

Lunin, V. Yu. (1988). *Use of the information on electron density distribution in macromolecules. Acta Cryst.* A**44**, 144–150.

Lunin, V. Yu. & Skovoroda, T. P. (1991). *Frequency-restrained structure-factor refinement. I. Histogram simulation. Acta Cryst.* A**47**, 45–52.

Main, P. (1979). *A theoretical comparison of the β, γ' and $2F_o - F_c$ syntheses. Acta Cryst.* A**35**, 779–785.

Main, P. (1990*a*). *A formula for electron density histograms for equal-atom structures. Acta Cryst.* A**46**, 507–509.

Main, P. (1990*b*). *The use of Sayre's equation with constraints for the direct determination of phases. Acta Cryst.* A**46**, 372–377.

Main, P. & Rossmann, M. G. (1966). *Relationships among structure factors due to identical molecules in different crystallographic environments. Acta Cryst.* **21**, 67–72.

Matthews, B. W. (1968). *Solvent content of protein crystals. J. Mol. Biol.* **33**, 491–497.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *On the application of the minimal principle to solve unknown structures. Science*, **259**, 1430–1433.

Navaza, J. (1994). *AMoRe: an automated package for molecular replacement. Acta Cryst.* A**50**, 157–163.

Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *wARP: improvement and extension of crystallographic phases by weighted averaging of multiple-refined dummy atomic models. Acta Cryst.* D**53**, 448–455.

Podjarny, A. D. & Yonath, A. (1977). *Use of matrix direct methods for low-resolution phase extension for tRNA. Acta Cryst.* A**33**, 655–661.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. & Schierbeek, A. J. (1988). *A phased translation function. J. Appl. Cryst.* **21**, 490–495.

Reynolds, R. A., Remington, S. J., Weaver, L. H., Fisher, R. G., Anderson, W. F., Ammon, H. L. & Matthews, B. W. (1985). *Structure of a serine protease from rat mast cells determined from twinned crystals by isomorphous and molecular replacement. Acta Cryst.* B**41**, 139–147.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst.* **15**, 24–31.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *Molecular replacement real-space averaging. J. Appl. Cryst.* **25**, 166–180.

Sayre, D. (1952). *The squaring method: a new method for phase determination. Acta Cryst.* **5**, 60–65.

Sayre, D. (1972). *On least-squares refinement of the phases of crystallographic structure factors. Acta Cryst.* A**28**, 210–212.

Sayre, D. (1974). *Least-squares phase refinement. II. High-resolution phasing of a small protein. Acta Cryst.* A**30**, 180–184.

Schevitz, R. W., Podjarny, A. D., Zwick, M., Hughes, J. J. & Sigler, P. B. (1981). *Improving and extending the phases of medium- and low-resolution macromolecular structure factors by density modification. Acta Cryst.* A**37**, 669–677.

Schuller, D. J. (1996). *MAGICSQUASH: more versatile non-crystallographic averaging with multiple constraints*. *Acta Cryst.* D**52**, 425–434.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst.* **12**, 813–815.

Swanson, S. M. (1994). *Core tracing: depicting connections between features in electron density. Acta Cryst.* D**50**, 695–708.

Terwilliger, T. C. (1999). *Reciprocal-space solvent flattening*. *Acta Cryst.* D**55**, 1863–1871.

Terwilliger, T. C. (2004). *Using prime-and-switch phasing to reduce model bias in molecular replacement*. *Acta Cryst.* D**60**, 2144–2149.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Ab initio phase determination for viruses with high symmetry: a feasibility study. Acta Cryst.* A**48**, 293–301.

Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE*. *Acta Cryst.* D**63**, 1069–1074.

Vellieux, F. M. D. (1998). *A comparison of two algorithms for electron-density map improvement by introduction of atomicity: skeletonization, and map sorting followed by refinement. Acta Cryst.* D**54**, 81–85.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification. J. Appl. Cryst.* **28**, 347–351.

Wang, B. C. (1985). *Resolution of phase ambiguity in macromolecular crystallography*. In *Diffraction Methods for Biological Macromolecules*, edited by H. W. Wyckoff, C. H. W. Hirs & S. N. Timasheff, Vol. 115, pp. 90–113. Orlando: Academic Press.

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Applications of the minimal principle to peptide structures. Acta Cryst.* D**49**, 179–181.

Wigley, D. B., Roper, D. I. & Cooper, R. A. (1989). *Preliminary crystallographic analysis of 5-carboxymethyl-2-hydroxymuconate isomerase from Escherichia coli. J. Mol. Biol.* **210**, 881–882.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.

Wilson, C. & Agard, D. A. (1993). *PRISM: automated crystallographic phase refinement by iterative skeletonization. Acta Cryst.* A**49**, 97–104.

Woolfson, M. M. (1987). *Direct methods – from birth to maturity. Acta Cryst.* A**43**, 593–612.

Zhang, K. Y. J. (1993). *SQUASH – combining constraints for macromolecular phase refinement and extension. Acta Cryst.* D**49**, 213–222.

Zhang, K. Y. J., Cowtan, K. D. & Main, P. (1997). *Combining constraints for electron density modification.* In *Macromolecular Crystallography*, edited by C. W. Carter & R. M. Sweet, Vol. 277, pp. 53–64. New York: Academic Press.

Zhang, K. Y. J. & Main, P. (1988). *Histogram matching as a density modification technique for phase refinement and extension of protein molecules.* In *Improving Protein Phases*, edited by S. Bailey, E. Dodson & S. Phillips. Report DL/SCI/R26, pp. 57–64. Warrington: Daresbury Laboratory.

Zhang, K. Y. J. & Main, P. (1990*a*). *Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. Acta Cryst.* A**46**, 41–46.

Zhang, K. Y. J. & Main, P. (1990*b*). *The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. Acta Cryst.* A**46**, 377–381.

International Tables for Crystallography (2012). Vol. F, Chapter 15.2, pp. 401–406.

Chapter 15.2. Model phases: probabilities, bias and maps

R. J. Read

15.2.1. Introduction

The intensities of X-ray diffraction spots measured from a crystal give us only the amplitudes of the diffracted waves. To reconstruct a map of the electron density in the crystal, the unmeasured phase information is also required. In fact, the phases are much more important to the appearance of the map than the measured amplitudes. When phases are supplied by an atomic model, therefore, some degree of model bias is inevitable.

The optimal use of model phase information requires an estimate of its reliability, specifically the probability that various values of the phase angle are true. Such a probability distribution can be derived, starting first with the relationship between the structure factor (amplitude and phase) of the model and that of the true crystal structure. The phase probability distribution can then be obtained from this and used, for instance, to provide a figure-of-merit weighting that minimizes the r.m.s. error from the true electron density.

Even with figure-of-merit weighting, model-phased electron density is biased towards the model. The systematic bias component of model-phased map coefficients can be predicted, allowing the derivation of map coefficients that give electron-density maps with reduced model bias. With the help of a few simple assumptions, a correction for bias can also be made when different sources of phase information are combined.

Finally, the refinement of a model against the observed amplitudes allows a certain amount of overfitting of the data, which leads to an extra 'refinement bias'. Fortunately, the use of appropriate refinement strategies, including maximum-likelihood targets, can reduce the severity of this problem.

15.2.2. Model bias: importance of phase

Dramatic illustrations of the importance of the phase have been published. For instance, Ramachandran & Srinivasan (1961) calculated an electron-density map using phases from one structure and amplitudes from another. In this map there were peaks at the positions of the atoms in the structure that contributed the phase information, but not in the structure that contributed the amplitudes. Similar calculations with two-dimensional Fourier transforms of photographs (Oppenheim & Lim, 1981; Read, 1997) showed that the phases of one completely overwhelmed the amplitudes of the other.

These examples, though dramatic, are not completely representative of the normal situation, where the structure contributing the phases is partially or even nearly correct. Nonetheless, model phases always contribute bias, so that the resulting map tends to bear too close a resemblance to the model.

15.2.2.1. Parseval's theorem

The importance of the phase can be understood most easily in terms of Parseval's theorem, a result that is important to the understanding of many aspects of the Fourier transform and its use in crystallography. Parseval's theorem states that the mean-square value of the variable on one side of a Fourier transform is proportional to the mean-square value of the variable on the other side. Since the Fourier transform is additive, Parseval's theorem also applies to sums or differences.

If ρ_1 and ρ_2 are, for instance, the true electron density and the electron density of the model, respectively, Parseval's theorem tells us that the r.m.s. error in the electron density is proportional to the r.m.s. error in the structure factor. (The structure-factor error is a vector error in the complex plane.)

$$\langle \rho^2 \rangle = (1/V^2) \sum_{\text{all } \mathbf{h}} |\mathbf{F}(\mathbf{h})|^2,$$

$$\left\langle (\rho_1 - \rho_2)^2 \right\rangle = (1/V^2) \sum_{\text{all } \mathbf{h}} |\mathbf{F}_1(\mathbf{h}) - \mathbf{F}_2(\mathbf{h})|^2.$$

This understanding of error in electron-density maps explains why the phase is much more important than the amplitude in determining the appearance of an electron-density map. As illustrated in Fig. 15.2.2.1, a random choice of phase (from a uniform distribution of all possible phases) will generally give a larger error in the complex plane than a random choice of amplitude [from a Wilson (1949) distribution of amplitudes].

15.2.3. Structure-factor probability relationships

To use model phase information optimally, the probability distribution for the true phase (or, equivalently, the distribution of the error in the model phase) needs to be known. Such a distribution can be derived by first working out the probability distribution for the true structure factor (or the distribution of the vector difference between the model and true structure factors). Then the phase probability distribution is obtained by fixing the known value of the structure-factor amplitude and renormalizing.

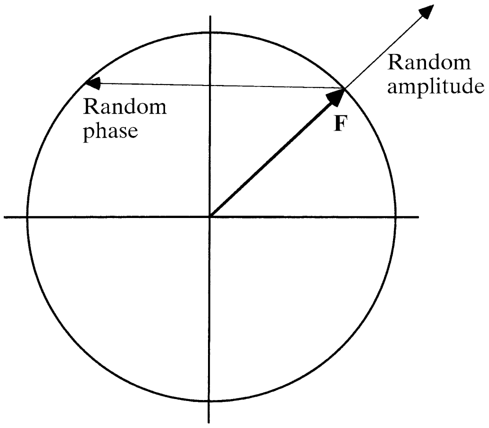


Figure 15.2.2.1
Schematic illustration of the relative errors introduced by a random choice of phase or a random choice of amplitude. The example has been constructed to represent the r.m.s. errors introduced by randomization (computed by averages over the Wilson distribution). Phase randomization will introduce r.m.s. errors of $(2)^{1/2}$ ($\simeq 1.41$) times the r.m.s. structure-factor amplitude $|\mathbf{F}|$. By comparison, map coefficients weighted by figures of merit of zero would have r.m.s. errors equal to the r.m.s. $|\mathbf{F}|$, so a featureless map would be more accurate than a random-phase map. Amplitude randomization will introduce r.m.s. errors of $[(4-\pi)/2]^{1/2}$ ($\simeq 0.66$) times the r.m.s. $|\mathbf{F}|$, so a map computed with random amplitudes will be closer to the true map than a featureless map.

Copyright © 2012 International Union of Crystallography

A number of related structure-factor distributions have been derived, differing in the amount of information available about the structure and in the assumed form of errors in the model. These range from the Wilson distribution, which applies when none of the atomic positions is known, to a distribution that applies when there are a variety of sources of error in an atomic model.

15.2.3.1. Wilson and Sim structure-factor distributions in P1

For the Wilson distribution (Wilson, 1949), it is assumed that the atoms in a crystal structure in space group $P1$ are scattered randomly and independently through the unit cell. In fact, it is sufficient to make the much less restrictive assumption that the atoms are placed randomly with respect to the Bragg planes defined by the Miller indices. The assumption of independence is somewhat more problematic, since there are restrictions on the distances between atoms, large volumes of protein crystals are occupied by disordered solvent and many protein crystals display noncrystallographic symmetry; as discussed elsewhere (Vellieux & Read, 1997; Kleywegt & Read, 1997), the resulting relationships among structure factors are exploited implicitly in averaging and solvent-flattening procedures. The higher-order relationships among structure factors are used explicitly in direct methods for solving small-molecule structures and are being developed for use in protein structures (Bricogne, 1993). For the purposes of simpler relationships between the calculated and true structure factors for a single *hkl*, however, the lack of complete independence does not seem to create serious problems.

When atoms are placed randomly relative to the Bragg planes, the contribution of each atom to the structure factor will have a phase varying randomly from 0 to 2π. The overall structure factor can then be considered to be the result of a random walk in the complex plane, which can be treated as an application of the central limit theorem. The structure factor is the sum of the independent atomic scattering contributions, each of which has a probability distribution defined as a circle in the complex plane centred on the origin, with a radius of f_j. The centroid of this atomic distribution is at the origin, and the variance for each of the real and imaginary parts is $\frac{1}{2}f_j^2$. The probability distribution of the structure factor that is the sum of these contributions is a two-dimensional Gaussian, the product of the one-dimensional Gaussians for the real and imaginary parts. Because the variances are equal in the real and imaginary directions, it can be simplified, as shown below, and expressed in terms of a single distribution parameter, Σ_N. Alternatively, and more simply, the structure-factor probability distribution can be considered as a complex normal distribution, arising from application of the central limit theorem to complex variables (Read, 2003).

$$\mathbf{F} = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\cdot\mathbf{x}_j) = A + iB; \quad \langle A\rangle = \langle B\rangle = 0;$$

$$\sigma^2(A) = \sigma^2(B) = \tfrac{1}{2}\sum_{j=1}^{N} f_j^2 = \tfrac{1}{2}\Sigma_N, \text{ so}$$

$$\begin{aligned} p(A) &= [1/(\pi\Sigma_N)^{1/2}]\exp\left(-A^2/\Sigma_N\right),\\ p(B) &= [1/(\pi\Sigma_N)^{1/2}]\exp\left(-B^2/\Sigma_N\right),\\ p(\mathbf{F}) = p(A,B) &= (1/\pi\Sigma_N)\exp\left(-|\mathbf{F}|^2/\Sigma_N\right). \end{aligned}$$

The Sim distribution (Sim, 1959), which is relevant when the positions of some of the atoms are known, has a very similar basis, except that the structure factor is now considered to arise from a random walk starting from the position of the structure factor corresponding to the known part, $\mathbf{F}_P$. Atoms with known positions do not contribute to the variance, while each of the atoms with unknown positions (the 'Q' atoms) contributes $\frac{1}{2}f_j^2$ to each of the real and imaginary parts, as in the Wilson distribution. The distribution parameter in this case is referred to as Σ_Q. The Sim distribution is a conditional probability distribution, depending on the value of $\mathbf{F}_P$,

$$p(\mathbf{F};\mathbf{F}_P) = (1/\pi\Sigma_Q)\exp\left(-|\mathbf{F}-\mathbf{F}_P|^2/\Sigma_Q\right).$$

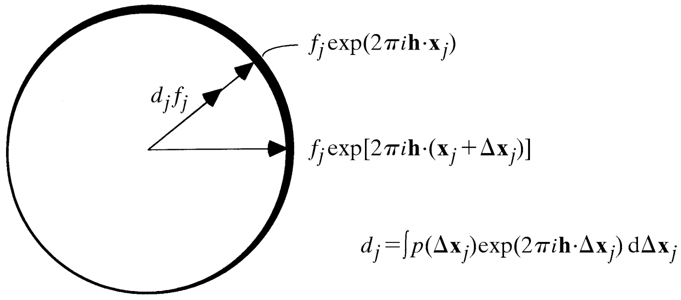

$f_j\exp(2\pi i\mathbf{h}\cdot\mathbf{x}_j)$
$d_j f_j$
$f_j\exp[2\pi i\mathbf{h}\cdot(\mathbf{x}_j+\Delta\mathbf{x}_j)]$
$d_j=\int p(\Delta\mathbf{x}_j)\exp(2\pi i\mathbf{h}\cdot\Delta\mathbf{x}_j)\,\mathrm{d}\Delta\mathbf{x}_j$

Figure 15.2.3.1
Centroid of the structure-factor contribution from a single atom. The probability of a phase for the contribution is indicated by the thickness of the line.

The Wilson (1949) and Woolfson (1956) distributions for space group $P\bar{1}$ are obtained similarly, except that the random walks are along a line and the resulting Gaussian distributions are one-dimensional. (The Woolfson distribution is the centric equivalent of the Sim distribution.) For more complicated space groups, it is reasonable to assume that acentric reflections follow the $P1$ distribution and that centric reflections follow the $P\bar{1}$ distribution. However, for any zone of the reciprocal lattice in which symmetry-related atoms are constrained to scatter in phase, the variances must be multiplied by the expected intensity factor, ε, for the zone, because the symmetry-related contributions are no longer independent.

15.2.3.2. Probability distributions for variable coordinate errors

In the Sim distribution, an atom is considered to be either exactly known or completely unknown in its position. These are extreme cases, since there will normally be varying degrees of uncertainty in the positions of various atoms in a model. The treatment can be generalized by allowing a probability distribution of coordinate errors for each atom. In this case, the centroid for the individual atomic contribution to the structure factor will no longer be obtained by multiplying by either zero or one. Averaged over the circle corresponding to possible phase errors, the centroid will generally be reduced in magnitude, as illustrated in Fig. 15.2.3.1. In fact, averaging to obtain the centroid is equivalent to weighting the atomic scattering contribution by the Fourier transform of the coordinate-error probability distribution, d_j. By the convolution theorem, this in turn is equivalent to convoluting the atomic density with the coordinate-error distribution. Intuitively, the atom is smeared over all of its possible positions. The weighting factor, d_j, is thus analogous to the thermal-motion term in the structure-factor expression.

The variances for the individual atomic contributions will differ in magnitude, but if there are a sufficient number of independent sources of error, we can invoke the central limit theorem again and assume that the probability distribution for the structure factor will be a Gaussian centred on $\sum d_j f_j \exp\left(2\pi i\mathbf{h}\cdot\mathbf{x}_j\right)$. If the coordinate-error distribution is Gaussian, and if each atom in the model is subject to the same

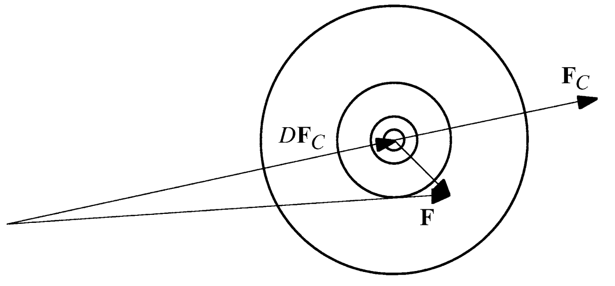


Figure 15.2.3.2
Schematic illustration of the general structure-factor distribution, relevant in the case of any set of independent random errors in the atomic model.

errors, the resulting structure-factor probability distribution is the Luzzati (1952) distribution. In this special case, $d_j = D$ for all atoms, where D is the Fourier transform of a Gaussian and behaves like the application of an overall B factor.

15.2.3.3. *General treatment of the structure-factor distribution*

The Wilson, Sim, Luzzati and variable-error distributions have very similar forms, because they are all Gaussians arising from the application of the central limit theorem. The central limit theorem is valid under many circumstances; even when there are errors in position, scattering factor and B factor, as well as missing atoms, a similar distribution still applies. As long as these sources of error are independent, the true structure factor will have a Gaussian distribution centred on $D\mathbf{F}_C$ (Fig. 15.2.3.2), where D now includes effects of all sources of error, as well as compensating for errors in the overall scale and B factor (Read, 1990).

$$p(\mathbf{F}; \mathbf{F}_C) = (1/\pi\varepsilon\sigma_\Delta^2)\exp\left(-|\mathbf{F} - D\mathbf{F}_C|^2/\varepsilon\sigma_\Delta^2\right)$$

in the acentric case, where $\sigma_\Delta^2 = \Sigma_N - D^2\Sigma_P$, ε is the expected intensity factor and Σ_P is the Wilson distribution parameter for the model.

For centric reflections, the scattering differences are distributed along a line, so the probability distribution is a one-dimensional Gaussian.

$$p(\mathbf{F}; \mathbf{F}_C) = [1/(2\pi\varepsilon\sigma_\Delta^2)^{1/2}]\exp\left(-|\mathbf{F} - D\mathbf{F}_C|^2/2\varepsilon\sigma_\Delta^2\right).$$

15.2.3.4. *Estimating σ_A*

Srinivasan (1966) showed that the Sim and Luzzati distributions could be combined into a single distribution that had a particularly elegant form when expressed in terms of normalized structure factors, or E values. This functional form still applies to the general distribution that reflects a variety of sources of error; the only difference is the interpretation placed on the parameters (Read, 1990). If $\mathbf{F}$ and $\mathbf{F}_C$ are replaced by the corresponding E values, a parameter σ_A plays the role of D, and σ_Δ^2 reduces to $(1 - \sigma_A^2)$. [The parameter σ_A is equivalent to D after correction for model completeness; $\sigma_A = D(\Sigma_P/\Sigma_N)^{1/2}$.] When the structure factors are normalized, overall scale and B-factor effects are also eliminated. The parameter σ_A that characterizes this probability distribution varies as a function of resolution. It must be deduced from the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$, since the phase (thus the phase difference) is unknown.

A general approach to estimating parameters for probability distributions is to maximize a likelihood function. The likelihood function is the overall joint probability of making the entire set of observations, which is a function of the desired parameters. The parameters that maximize the probability of making the set of observations are the most consistent with the data. The idea of using maximum likelihood to estimate model phase errors was introduced by Lunin & Urzhumtsev (1984), who gave a treatment that was valid for space group $P1$. In a more general treatment that applies to higher-symmetry space groups, allowance is made for the statistical effects of crystal symmetry (centric zones and differing expected intensity factors) (Read, 1986).

The σ_A values are estimated by maximizing the joint probability of making the set of observations of $|\mathbf{F}_O|$. If the structure factors are all assumed to be independent, the joint probability distribution is the product of all the individual distributions. The assumption of independence is not completely justified in theory, but the results are fairly accurate in practice.

$$L = \prod_{\mathbf{h}} p(|\mathbf{F}_O|; |\mathbf{F}_C|).$$

The required probability distribution, $p(|\mathbf{F}_O|; |\mathbf{F}_C|)$, is derived from $p(\mathbf{F}_O; \mathbf{F}_C)$ by integrating over all possible phase differences. The form of this distribution, which is given in other publications (Read, 1986, 1990), differs for centric and acentric reflections. (It is important to note that although the distributions for structure factors are Gaussian, the distributions for amplitudes obtained by integrating out the phase are not.) It is more convenient to deal with a sum than a product, so the log likelihood function is maximized instead. In the program *SIGMAA*, reciprocal space is divided into spherical shells, and a value of the parameter σ_A is refined for each resolution shell. Details of the algorithm are given elsewhere (Read, 1986).

The resolution shells must be thick enough to contain several hundred to a thousand reflections each, in order to provide σ_A estimates with a sufficiently small statistical error. A larger number of shells (fewer reflections per shell) can be used for refined structures, since estimates of σ_A become more precise as the true value approaches 1. If there are sufficient reflections per shell, the estimates will vary smoothly with resolution. As discussed below, the smooth variation with resolution can also be exploited through a restraint that allows σ_A values to be estimated from fewer reflections.

15.2.3.5. *Accounting for measurement errors*

For unrefined models, the disagreement between the true and calculated structure factors is dominated by the effect of defects in the model, so that the effect of errors in measuring the observed intensities or amplitudes can be neglected (Read, 1986). As the model improves, the effect of measurement error becomes relevant. Three approaches have been used to account for measurement error. In one approach, first proposed by Green (1979) in the context of isomorphous replacement, measurement error is treated as a complex error, and the variance term, σ, is increased accordingly (Murshudov *et al.*, 1997; Bricogne & Irwin, 1996). This has the virtue of simplicity and it works well. Another approach is to use a Gaussian approximation to the likelihood function expressed in terms of amplitudes, and then increase the variance of this Gaussian (Pannu & Read, 1996). Finally, the convolution of the likelihood function (expressed in intensities) and a Gaussian measurement error can be evaluated by using a series approximation (Pannu & Read, 1996). This has the benefit that it allows for the presence of negative net intensity measurements, but it is more complicated to implement and is not widely used.

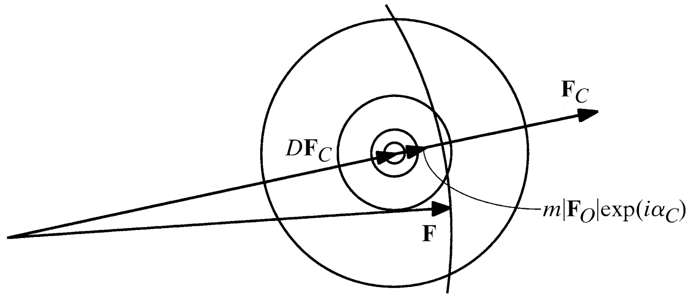


Figure 15.2.4.1
Figure-of-merit weighted model-phased structure factor, obtained as the probability-weighted average over all possible phases.

15.2.4. Figure-of-merit weighting for model phases

Blow & Crick (1959) and Sim (1959) showed that the electron-density map with the least r.m.s. error is calculated from centroid structure factors. This conclusion follows from Parseval's theorem, because the centroid structure factor (its probability-weighted average value or expected value) minimizes the r.m.s. error of the structure factor. Since the structure-factor distribution $p(\mathbf{F};\,\mathbf{F}_C)$ is symmetrical about $\mathbf{F}_C$, the expected value of $\mathbf{F}$ will have the same phase as $\mathbf{F}_C$, but the averaging around the phase circle will reduce its magnitude if there is any uncertainty in the phase value (Fig. 15.2.4.1). We treat the reduction in magnitude by applying a weighting factor called the figure of merit, m, which is equivalent to the expected value of the cosine of the phase error.

15.2.5. Map coefficients to reduce model bias

15.2.5.1. Model bias in figure-of-merit weighted maps

A figure-of-merit weighted map, calculated with coefficients $m|\mathbf{F}_O|\exp(i\alpha_C)$, has the least r.m.s. error from the true map. According to the normal statistical (minimum variance) criteria, then, it is the best map. However, such a map will suffer from model bias; if its purpose is to allow the detection and repair of errors in the model, this is a serious qualitative defect. Fortunately, it is possible to predict the systematic errors leading to model bias and to make some correction for them.

Main (1979) dealt with this problem in the case of a perfect partial structure. Since the relationships among structure factors are the same in the general case of a partial structure with various errors, once $D\mathbf{F}_C$ is substituted for $\mathbf{F}_C$, all that is required to apply Main's results more generally is a change of variables (Read, 1986, 1990).

In Main's approach, the cosine law is used to introduce the cosine of the phase error, which is converted into a figure of merit by taking expected values. Some manipulations allow us to solve for the figure-of-merit weighted map coefficient, which is approximated as a linear combination of the true structure factor and the model structure factor (Main, 1979; Read, 1986). Finally, we can solve for an approximation to the true structure factor, giving map coefficients from which the systematic model bias component has been removed.

$$m|\mathbf{F}_O|\exp(i\alpha_C) = F/2 + D\mathbf{F}_C/2 + \text{noise terms},$$
$$F \simeq (2m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C).$$

A similar analysis for centric structure factors shows that there is no systematic model bias in figure-of-merit weighted map coefficients, so no bias correction is needed in the centric case.

15.2.5.2. Model bias in combined phase maps

When model phase information is combined with, for instance, multiple isomorphous replacement (MIR) phase information, there will still be model bias in the acentric map coefficients, to the extent that the model influences the final phases. However, it is inappropriate to continue using the same map coefficients to reduce model bias, because some phases could be determined almost completely by the MIR phase information. It makes much more sense to have map coefficients that reduce to the coefficients appropriate for either model or MIR phases, in extreme cases where there is only one source of phase information, and that vary smoothly between those extremes.

Map coefficients that satisfy these criteria (even if they are not rigorously derived) are implemented in the program *SIGMAA*. The resulting maps are reasonably successful in reducing model bias. Two assumptions are made: (1) the model-bias component in the figure-of-merit weighted map coefficient, $m_{\rm com}|\mathbf{F}_O|\exp(i\alpha_{\rm com})$, is proportional to the influence that the model phase has had on the combined phase; and (2) the relative influence of a source of phase information can be measured by the information content, H (Guiasu, 1977), of the phase probability distribution. The first assumption corresponds to the idea that the figure-of-merit weighted map coefficient is a linear combination of the MIR and model phase cases.

$$\begin{array}{lll} \text{MIR:} & m_{\rm MIR}|\mathbf{F}_O|\exp(i\alpha_{\rm MIR}) & \simeq \mathbf{F} \\ \text{Model:} & m_C|\mathbf{F}_O|\exp(i\alpha_C) & \simeq \mathbf{F}/2 + D\mathbf{F}_C/2 \\ \text{Combined:} & m_{\rm com}|\mathbf{F}_O|\exp(i\alpha_{\rm com}) & \simeq [1-(w/2)]\mathbf{F} + (w/2)D\mathbf{F}_C, \end{array}$$

where

$$w = H_C/(H_C + H_{\rm MIR})$$

and

$$H = \int\limits_0^{2\pi} p(\alpha)\ln\frac{p(\alpha)}{p_0(\alpha)}\,{\rm d}\alpha; \qquad p_0(\alpha) = \frac{1}{2\pi}.$$

Solving for an approximation to the true $\mathbf{F}$ gives the following expression, which can be seen to reduce appropriately when w is 0 (no model influence) or 1 (no MIR influence):

$$\mathbf{F} \simeq \frac{2m|\mathbf{F}_O|\exp(i\alpha_{\rm com}) - wD\mathbf{F}_C}{2-w}.$$

15.2.6. Estimation of overall coordinate error

In principle, since the distribution of observed and calculated amplitudes is determined largely by the coordinate errors of the model, one can determine whether a particular coordinate-error distribution is consistent with the amplitudes. Unfortunately, it turns out that the coordinate errors cannot be deduced unambiguously, because many distributions of coordinate errors are consistent with a particular distribution of amplitudes (Read, 1990).

If the simplifying assumption is made that all the atoms are subject to a single error distribution, then the parameter D (and thus the related parameter σ_A) varies with resolution as the Fourier transform of the error distribution, as discussed above. Two related methods to estimate overall coordinate error are based on the even more specific assumption that the coordinate-error distribution is Gaussian: the Luzzati plot (Luzzati, 1952) and the σ_A plot (Read, 1986). Unfortunately, the central assumption is not justified; atoms that scatter more strongly (heavier atoms or atoms with lower B factors) tend to have smaller coordinate errors than weakly scattering atoms. The proportion of the structure factor contributed by well ordered atoms increases at high resolution, so that the structure factors agree better at high resolution than if there were a single error distribution.

It is often stated, optimistically, that the Luzzati plot provides an upper bound to the coordinate error, because the observation errors in $|\mathbf{F}_O|$ have been ignored. This is misleading, because there are other effects that cause the Luzzati and σ_A plots to give underestimates (Read, 1990). Chief among these are the correlation of errors and scattering power and the overfitting of the amplitudes in structure refinement (discussed below). These estimates of overall coordinate error should not be interpreted too literally; at best, they provide a comparative measure.

15.2.7. Difference-map coefficients

The computer program *SIGMAA* (Read, 1986) has been developed to implement the results described here. Apart from the two types of map coefficient discussed above, two types of difference-map coefficient can also be produced:

(1) Model-phased difference map: $(m|\mathbf{F}_O| - D|\mathbf{F}_C|)\exp(i\alpha_C)$;
(2) General difference map: $m_{\rm com}|\mathbf{F}_O|\exp(i\alpha_{\rm com}) - D\mathbf{F}_C$.

The general difference map, it should be noted, uses a vector difference between the figure-of-merit weighted combined phase coefficient (the 'best' estimate of the true structure factor) and the calculated structure factor. When additional phase information is available, it should provide a clearer picture of the errors in the model.

Similar algorithms have been implemented as part of modern refinement programs such as *Refmac* (Murshudov *et al.*, 1997) and *phenix.refine* (Afonine *et al.*, 2005).

15.2.8. Refinement bias

The structure-factor probabilities discussed above depend on the atoms having independent errors (or at least a sufficient number of groups of atoms having independent errors). Unfortunately, this assumption breaks down when a structure is refined against the observed diffraction data. Few protein crystals diffract to sufficiently high resolution to provide a large number of observations for every refinable parameter. The refinement problem is, therefore, not sufficiently overdetermined, so it is possible to overfit the data. If there is an error in the model that is outside the range of convergence of the refinement method, it is possible to introduce compensating errors in the rest of the structure to give a better, and misleading, agreement in the amplitudes. As a result, the phase accuracy (hence the weighting factors m and D) is overestimated, and model bias is poorly removed.

There is another interpretation to the problem of refinement bias. As Silva & Rossmann (1985) point out, minimizing the r.m.s. difference between the amplitudes $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$ is equivalent (by Parseval's theorem) to minimizing the difference between the model electron density and the density corresponding to the map coefficients $|\mathbf{F}_O|\exp(i\alpha_C)$; a lower residual is obtained either by making the model look more like the true structure, or by making the model-phased map look more like the model through the introduction of systematic phase errors.

A number of strategies are available to reduce the degree or impact of refinement bias. The overestimation of phase accuracy can be avoided by obtaining unbiased estimates of the σ_A values from the cross-validation data, which were originally introduced to compute $R_{\rm free}$ as an unbiased indicator of refinement progress (Brünger, 1992). Because of the high statistical error of σ_A estimates computed from small numbers of reflections, reliable values can only be obtained by exploiting the smoothness of the σ_A curve as a function of resolution. This can be achieved either by fitting a functional form (as done in *Refmac*; Murshudov *et al.*, 1997) or by adding a penalty to points that deviate from the line connecting their neighbours (as done in *CNS*; Brunger *et al.*, 1998).

If errors are suspected in certain parts of the structure, 'omit refinement' (in which the questionable parts are omitted from the model) can be a very effective way to eliminate refinement bias in those regions (James *et al.*, 1980; Hodel *et al.*, 1992).

If MIR or MAD (multiwavelength anomalous dispersion) phases are available, combined phase maps tend to suffer less from refinement bias, depending on the extent to which the experimental phases influence the combined phases. Finally, it is always a good idea to refer occasionally to the original MIR or MAD map, which cannot suffer at all from model bias or refinement bias.

This chapter is a revised version of a contribution to *Methods in Enzymology* (Read, 1997).

References

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsletter*, No. 42, contribution 8.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Bricogne, G. (1993). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.

Bricogne, G. & Irwin, J. (1996). In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–474.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR System: a new software suite for macromolecular structure determination*. *Acta Cryst.* D**54**, 905–921.

Green, E. A. (1979). *A new statistical model for describing errors in isomorphous replacement data: the case of one derivative*. *Acta Cryst.* A**35**, 351–359.

Guiasu, S. (1977). *Information Theory with Applications.* London: McGraw-Hill.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

James, M. N. G., Sielecki, A. R., Brayer, G. D., Delbaere, L. T. J. & Bauer, C.-A. (1980). *Structures of product and inhibitor complexes of Streptomyces griseus protease A at 1.8 Å resolution – a model for serine protease catalysis. J. Mol. Biol.* **144**, 43–88.

Kleywegt, G. J. & Read, R. J. (1997). *Ways & means: not your average density*. *Structure*, **5**, 1557–1569.

Lunin, V. Yu. & Urzhumtsev, A. G. (1984). *Improvement of protein phases by coarse model modification. Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst.* **5**, 802–810.

Main, P. (1979). *A theoretical comparison of the* β, γ' *and* $2F_o - F_c$ *syntheses. Acta Cryst.* A**35**, 779–785.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Oppenheim, A. V. & Lim, J. S. (1981). *The importance of phase in signals. Proc. IEEE*, **69**, 529–541.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Ramachandran, G. N. & Srinivasan, R. (1961). *An apparent paradox in crystal structure analysis. Nature (London)*, **190**, 159–161.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.

Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol.* **277**, 110–128.

Read, R. J. (2003). *New ways of looking at experimental phasing. Acta Cryst.* D**59**, 1891–1902.

Silva, A. M. & Rossmann, M. G. (1985). *The refinement of southern bean mosaic virus in reciprocal space. Acta Cryst.* B**41**, 147–157.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst.* **12**, 813–815.

Srinivasan, R. (1966). *Weighting functions for use in the early stages of structure analysis when a part of the structure is known. Acta Cryst.* **20**, 143–144.

Vellieux, F. M. D. & Read, R. J. (1997). *Non-crystallographic symmetry averaging in phase refinement and extension. Methods Enzymol.* **277**, 18–53.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.

Woolfson, M. M. (1956). *An improvement of the 'heavy-atom' method of solving crystal structures. Acta Cryst.* **9**, 804–810.

International Tables for Crystallography (2012). Vol. F, Chapter 15.3, pp. 407–412.

Chapter 15.3. *DM/DMMULTI* software for phase improvement by density modification

K. D. Cowtan, K. Y. J. Zhang and P. Main

15.3.1. Introduction

DM is an automated procedure for phase improvement by iterated density modification. It is used to obtain a set of improved phases and figures of merit, using as a starting point the observed diffraction amplitudes and some initial poor estimates for the phases and figures of merit. *DM* improves the phases through an alternate application of two processes: real-space electron-density modification and reciprocal-space phase combination. *DM* can perform solvent flattening, histogram matching, multi-resolution modification, averaging, skeletonization and Sayre refinement, as well as conventional or reflection-omit phase combination. Solvent and averaging masks may be input by the user or calculated automatically. Averaging operators may be refined within the program. Multiple averaging domains may be averaged using different operators.

DMMULTI is a modified version of the *DM* software that can perform density modification simultaneously across multiple crystal forms. The procedure is general, handling an arbitrary number of domains appearing in an arbitrary number of crystal forms. Initial phases may be provided for one or more crystal forms; however, improved phases are calculated in every crystal form.

DM and *DMMULTI* are distributed as a part of the *CCP4* suite of software for protein crystallography (Collaborative Computational Project, Number 4, 1994). The theoretical and algorithmic bases for the *DM* and *DMMULTI* software suites are reviewed in Chapter 15.1. In this chapter, some specific issues concerning the programs are described, including program operation, data preparation, choices of modes and code description.

15.3.2. Program operation

DM and *DMMULTI* are largely automatic; in order to perform a phase-improvement calculation only two tasks are required of the user:

(1) Provide the input data. These must include the reflection data and solvent content, and may also include averaging operators, solvent mask and averaging domain masks.
(2) Select the appropriate density modifications and the phase-combination mode to be used in the calculation.

DM and *DMMULTI* can run with the minimum input above, since the optimum choices for a whole range of parameters are set in the program defaults. For some special problems it may be useful to control the program behaviour in more detail; this is possible through a wide range of keywords to override the defaults. These are all detailed in the documentation supplied with the software.

15.3.3. Preparation of input data

Input data are provided by two routes: numerical parameters, such as solvent content and averaging operators, are included in the command file using appropriate keywords, whereas reflections and masks are referenced by giving their file names on the command line. In the simplest case; for example a solvent-flattening and histogram-matching calculation, all that is required is an initial reflection file and an estimate of the solvent content.

Use all available data: The reflection file must be in *CCP4* 'MTZ' format, and contain at least the structure-factor amplitudes, phase estimates and figures of merit. If the phase estimates are obtained from a homologous structure by molecular replacement, the figures of merit can be generated by the *SIGMAA* program (Read, 1986). When the phases are estimated using a single isomorphous derivative (SIR), it is recommended that Hendrickson–Lattman coefficients (Hendrickson & Lattman, 1970) are used to represent the phase estimate instead of the figure of merit. Hendrickson–Lattman coefficients can represent the bimodal distribution of the SIR phases, whereas the figure of merit can only represent the unimodal distribution of the average of two equally probable phase choices. It is recommended that a reflection file containing every possible reflection is used. The low-resolution data should be included since they provide a significant amount of information on the protein–solvent boundary. The high-resolution data without phase estimates should also be included since their phases can be estimated by *DM*. Phase extension can usually improve the original phases further compared to phase refinement only. Unobserved reflections are marked by a missing number flag. This is important for the preservation of the free-R reflections. It also enables *DM* to extrapolate missing reflections from density constraints and increases the phase improvement power.

The estimation of solvent content: The solvent content, $C_{\rm solv}$, can be obtained by various experimental methods, such as the solvent dehydration method and the deuterium exchange method (Matthews, 1974). It can also be estimated through

$$C_{\rm solv} = 1 - (NV_aML/V). \qquad (15.3.3.1)$$

Here, N is the total number of atoms, including hydrogen atoms, in one protein molecule. V_a is the average volume occupied by each atom, which is estimated to be approximately 10 Å^3 (Matthews, 1968). M is the number of molecules per asymmetric unit. L is the number of asymmetric units in the cell. V is the unit-cell volume. The correctly estimated solvent content should be entered in the program with the SOLC keyword, since this will be used not only to find the solvent–protein boundary but also to scale the input structure-factor amplitudes. If it is desirable to use a more conservative solvent mask in order to prevent clipping of protein densities, especially in the flexible loop regions, different solvent and protein fractions should be specified using the SOLMASK keyword.

Solvent mask: A solvent mask may be supplied; it may be used for the entire calculation or updated after several cycles. The solvent mask usually divides the cell into protein and solvent regions; however it is also possible to specify excluded regions which are unknown. If no solvent mask is supplied, it will be calculated by a modified Wang–Leslie procedure (Wang, 1985; Leslie, 1987) and updated as the phase-improvement calculation progresses.

Copyright © 2012 International Union of Crystallography

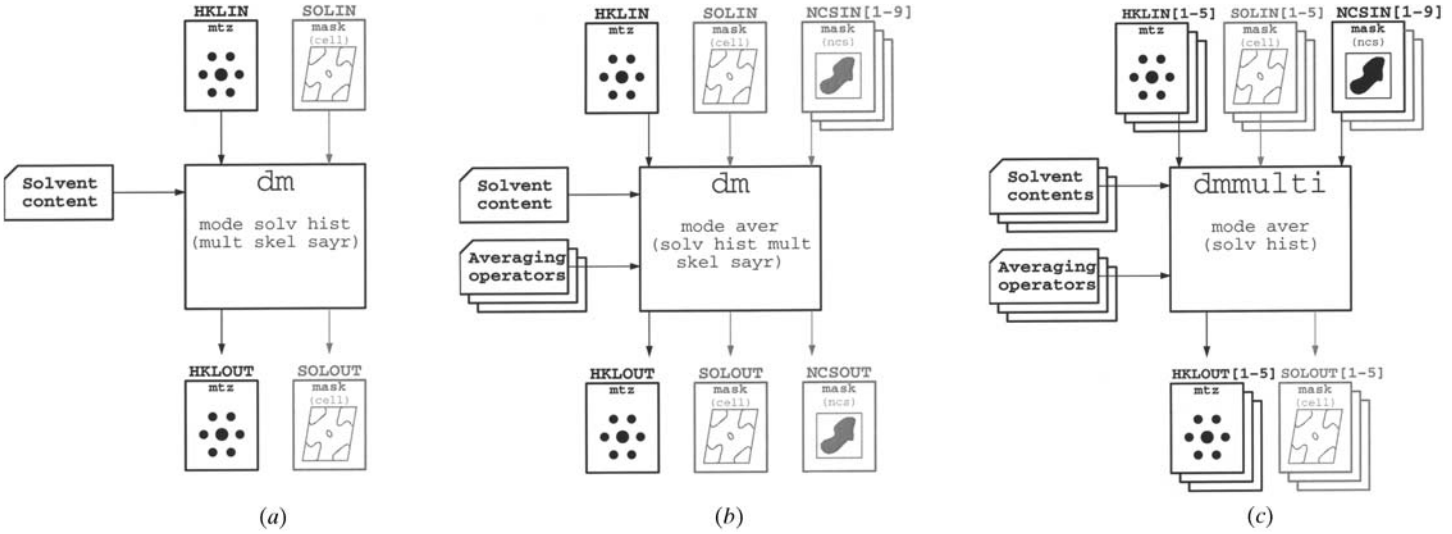


Figure 15.3.3.1

(*a*) Input and output data for a *DM* calculation with no averaging. Light outlines indicate optional information. (*b*) Input and output data for a *DM* averaging calculation: for a single averaging domain, the averaging mask may be calculated automatically. For multi-domain averaging, all domain masks must be given. (*c*) Input and output data for *DMMULTI*. An averaging mask (or masks, for multiple domains) must be provided.

Averaging operators: In an averaging calculation, the averaging operators must be supplied; these are typically obtained by rotation and translation searches using a program such as *AMoRe* (Navaza, 1994) or *X-PLOR* (Brünger, 1992). If the coordinates of several heavy atoms are known, they can be used to calculate the noncrystallographic symmetry (NCS) operators. If a partial model can be built into the density, structure-superposition programs, such as *LSQKAB* (Kabsch, 1976), can be used to obtain the rotation and translation matrices that relate different molecules in the asymmetric unit. This can also be achieved through the program *O* using the ‘lsq_explicit’ command (Jones *et al.*, 1991). The averaging operators can be further refined in *DM* by minimizing the residual between NCS related densities.

Averaging mask: An averaging mask may be supplied; this is distinct from the solvent mask, allowing for parts of the protein to remain unaveraged if required. If no averaging mask is supplied, the mask will be calculated by a local-correlation approach (Cowtan & Main, 1998; Vellieux *et al.*, 1995). If multiple domains are to be averaged with different averaging operators (Schuller, 1996), then one mask must be specified for each averaging domain. When averaging molecules related by improper NCS operations, the averaging mask must be in accord with the NCS operators provided. For example, if the supplied NCS matrix maps molecule A to molecule B, then the averaging mask must cover the volume occupied by molecule A rather than molecule B.

Multi-crystal averaging: In the case of a multi-crystal averaging calculation, one reflection file is provided for each crystal form (however, initial phases are not required in every crystal form), and one reflection file will be output for each crystal form containing the improved phases. One mask is required per averaging domain; thus, in general, only a single mask is required. This may be defined for any crystal form or in an arbitrary crystal space of its own. Averaging operators are then provided to map the mask into each of the crystal forms.

Solvent and averaging masks that are calculated within the program may be output for subsequent analysis. Refined averaging operators are also output. The input and output data for a simple *DM* calculation, a *DM* averaging calculation and a *DMMULTI* multi-crystal averaging calculation are shown in Figs. 15.3.3.1(*a*), (*b*) and (*c*), respectively.

15.3.4. Choice of modes

Two major choices have to be made in a *DM* run. They are the real-space density-modification modes and reciprocal-space phase-combination modes. Moreover, the phase-extension schemes can be selected if needed. This can also be left to the program, which uses its default automatic mode for phase extension. The choices of various modes are described in the following sections.

15.3.4.1. Density-modification modes

The following density-modification modes (specified by the MODE keyword) are provided by *DM*:

(1) Solvent flattening: This is the most common density-modification technique and is powerful for improving phases at fixed resolution, but weaker at extending phases to higher resolution. Its phasing power is highly dependent on the solvent content. Solvent flattening can be applied at comparatively low resolutions, down to around 5.0 Å.

(2) Histogram matching: This method is applied only to the density in the protein region. This method is weaker than solvent flattening for improving phases, but is much more powerful at extending phases to higher resolutions. This is due to a unique feature of histogram matching which uses a resolution-dependent target for phase improvement. The phasing power of histogram matching is inversely related to the solvent content. Therefore, histogram matching plays a more important role in phase improvement when the solvent content is low. Histogram matching works to as low as 4.0 Å, but does no harm below that. Histogram matching should probably be applied as a matter of course in any case where the structure is not dominated by a large proportion of heavy-metal atoms. Even in this case, histogram matching may be applied by defining a solvent mask with solvent, protein and excluded regions.

(3) Multi-resolution modification: This method controls the level of detail in the map as a function of resolution by applying histogram matching and solvent flattening at multiple resolutions. This strengthens phase improvement at fixed resolution, although it generally improves phase-extension calculations too.

(4) Noncrystallographic symmetry averaging: Averaging is one of the most powerful techniques available for improving phases and is applicable even at very low resolutions. In extreme cases, averaging may be used to achieve an *ab initio* structure solution (Chapman *et al.*, 1992; Tsao *et al.*, 1992). It should therefore be applied whenever it is present and the operators can be determined.
(5) Skeletonization: Iterative skeletonization is the process of tracing a 'skeleton' of connected densities through the map and then building a new map by filling density around this skeleton. The implementation in *DM* is adapted for use on poor maps, where it is sometimes but not always of use. To bring out side chains and missing loops, the *ARP* program (Lamzin & Wilson, 1997) is more suitable.
(6) Sayre's equation: This method is more widely used in small-molecule calculations, and is very powerful at better than 2.0 Å resolution and when there are no heavy atoms in the structure. However, its phasing power is lost quickly as resolution decreases beyond 2.0 Å. The calculation takes significantly longer than other density-modification modes.

The most commonly used modes are solvent flattening and histogram matching – these give a good first map in most cases. Recently, multi-resolution modification has been added to this list. Averaging is applied whenever possible. Skeletonization and Sayre's equation are generally only applied in special situations.

15.3.4.2. Phase-combination modes

Density-modification calculations are somewhat prone to producing grossly overestimated figures of merit (Cowtan & Main, 1996). Users should be aware of this. In general the phases and figures of merit produced by density-modification calculations should only be used for the calculation of weighted F_o maps. They should not be used for the calculation of difference maps or used in refinement or other calculations (the *REFMAC* program is an exception, containing a mechanism to deal with this form of bias). The use of $2F_o - F_c$-type maps should be avoided when the calculated phases are from density modification, since they are dependent on two assumptions, neither of which hold for density modification: that the current phases are very close to being correct and that the calculated amplitudes may only approach the observed values as the phase error approaches zero.

To limit the problems of overestimation, three phase-combination modes are provided (controlled by the COMBINE keyword):

(1) Free-Sim weighting: This is the simplest mode to use. Although convergence is weaker than the reflection-omit mode, the calculation never overshoots the best map. If there is averaging information, then convergence is much stronger and the phase-combination scheme is much less important. In addition, phase relationships in reciprocal space limit the effectiveness of the reflection-omit scheme. Therefore, the free-Sim weighting scheme should usually be used when there is averaging.
(2) Reflection-omit: The combination of a reciprocal-space omit procedure with *SIGMAA* phase combination (Read, 1986) leads to much better maps when applying solvent flattening and histogram matching. However, the omit calculation is computationally costly and introduces a small amount of noise into the maps, thus the phases can get worse if the calculation is run for too many cycles. A real-space free-*R* indicator (Abrahams & Leslie, 1996) is therefore used to stop the calculation at an appropriate point.
(3) Perturbation-γ correction: This new approach is an extension of the γ correction of Abrahams (1997) to arbitrary density-modification methods. The results are a good approximation to a perfect reflection-omit scheme and required considerably less computation. This is therefore the preferred mode for all calculations.

In the case of a molecular-replacement calculation or high noncrystallographic symmetry, it may be desirable only to weight the modified phases and not to recombine them back with the initial phases so that any initial bias may be overcome. In the case of high noncrystallographic symmetry, it may also be possible to restore missing reflections in both amplitude and phase. Options are available for both these situations.

15.3.4.3. Phase-extension schemes

When performing phase extension, the order in which the structure factors are included will affect the final accuracy of the extended phases. The phases obtained from previous cycles of phase extension will be included in the calculation of new phases for the unphased structure factors in the next cycle. A reflection with more accurately determined phases might enhance the phase-extension power of the original set of reflections, whereas a reflection with less accurately determined phases might corrupt the phase-extension power of the original set of reflections and make the phase extension deteriorate quickly. The factors that might affect phase extension are the structure-factor amplitudes, the resolution shell and the figure of merit. Based on the above considerations, the following phase-extension schemes are provided in *DM*:

(1) Extension by resolution shell: This performs phase extension in resolution steps, starting from the low-resolution data, and extends the phase to the high-resolution limit of the data or that specified by the user. Structure factors are related by the reciprocal-space density-modification function that is dominated by low-resolution terms, as shown by equation (15.1.3.2) and Fig. 15.1.3.1 in Chapter 15.1. This means that only structure factors in a small region of reciprocal space are related. Thus, when initial phases are only available at low resolution, phase extension is performed by inclusion of successive resolution shells. In the case of fourfold or higher NCS, this can allow extension to 2 Å starting from initial phasing at 6 Å or worse.
(2) Extension in structure-factor amplitude steps: In this mode, those reflections with larger amplitudes are added first, gradually extending to those reflections with smaller amplitudes in many steps. The contribution of a reflection to the electron density is proportional to the square of its structure-factor amplitude according to Parseval's theorem, as shown in equation (15.3.5.1). This favours the protocol of extending the stronger reflections first so that they can be more reliably estimated. These stronger reflections will be used to phase relatively weaker subsequent reflections.
(3) Extension in figure-of-merit step: To extend phases for those structure factors with experimentally measured, albeit less accurate, phases and figures of merit, the reflections can be added in order of their figure of merit, starting from the highest to the lowest. It is advantageous to use the more reliably estimated phases with higher figure of merit to phase those reflections with lower figure of merit. This can be useful

when working with initial phasing from multiwavelength anomalous diffraction (MAD) or molecular replacement (MR) sources.

(4) Automatic mode: This combines the previous three extension schemes. The program automatically works out the optimum combination of the above three schemes according to the density-modification mode, the phase-combination mode and the nature of the input reflection data. The automatic mode is the default and is the recommended mode of choice unless specific circumstances warrant a different choice.

(5) All reflection mode: One advantage of the reflection-omit and perturbation-γ methods is that the strength of extrapolation of a structure-factor amplitude is a good indicator of the reliability of its corresponding phase. As a result, a phase-extension scheme is unnecessary in reflection-omit calculations; all reflections may be included from the first cycle.

15.3.5. Code description

The program was designed to be run largely automatically with minimal user intervention. This is achieved by using extensive default settings and by automatic selection of options based on the data used. The program is also modular by design so that additional density-modification methods can be incorporated easily.

A simplified flow diagram for *DM* is shown in Fig. 15.3.5.1(*a*). When a reflection-omit calculation is performed, an additional loop is introduced, shown in Fig. 15.3.5.1(*b*). The Sayre's equation calculation adds another level of complexity, described in Zhang & Main (1990*b*). Skeletonization imposes the protein histogram and solvent flatness implicitly and so is performed, if necessary, every second or third cycle in place of solvent flattening and histogram matching. Simplified conceptual and actual flow diagrams for *DMMULTI* are shown in Figs. 15.3.5.2(*a*) and (*b*).

Many of the basic approaches used in *DM* and *DMMULTI* are described in Chapter 15.1. Some practical aspects of the application and combination of these approaches are described here.

15.3.5.1. Scaling

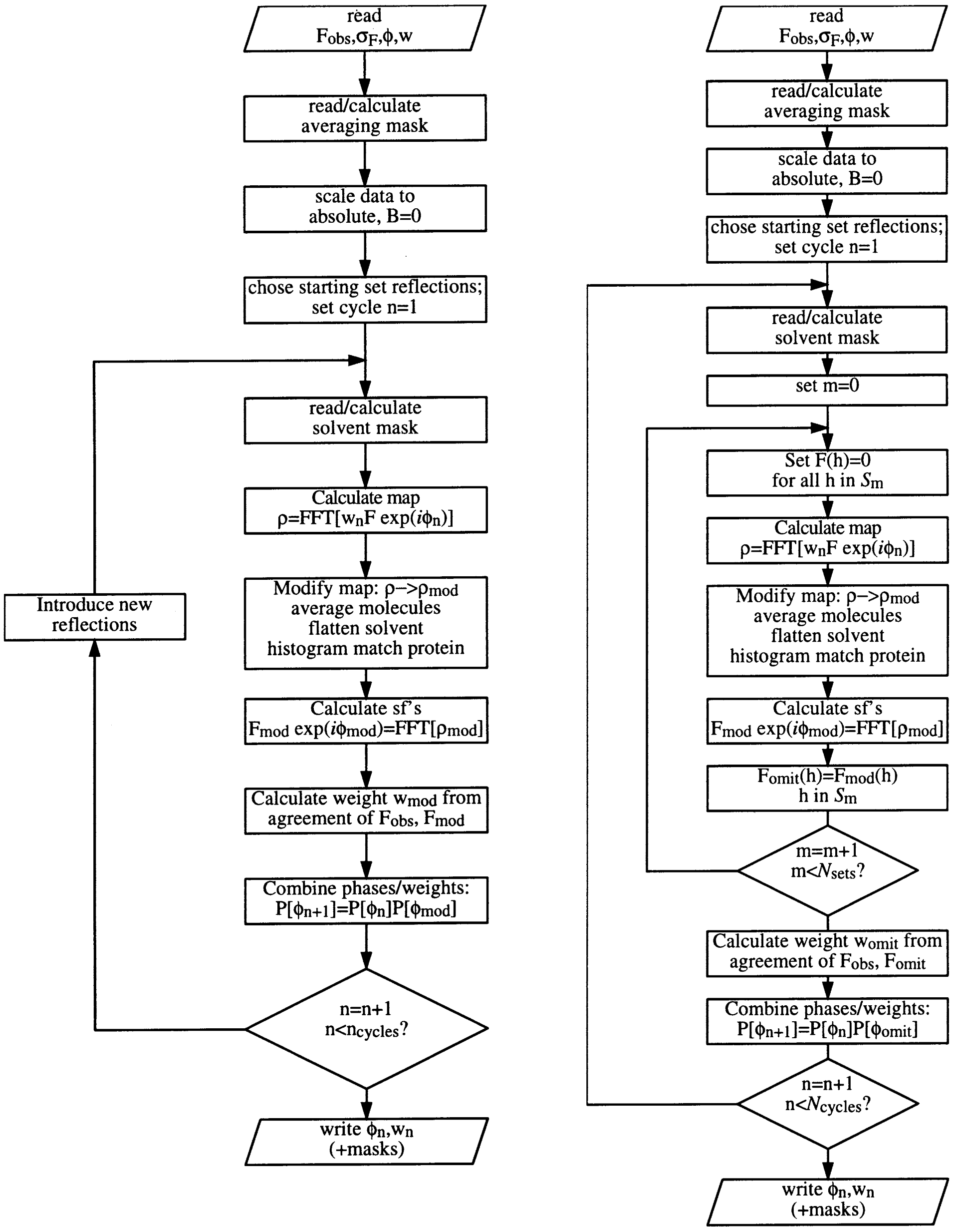
All forms of map modification are affected by the overall temperature factor of the data, and histogram matching in particular is critically dependent on the accurate determination of the scale factor. Wilson statistics have been found inadequate for scaling in this case, especially when the data resolution is worse than 3 Å, because of the dip in scattering below 5 Å.

More accurate estimates of the scale and temperature factors may be achieved by fitting the data to a semi-empirical scattering curve (Cowtan & Main, 1998). This curve is prepared using Parseval's theorem, which relates the sum of the intensities to the variance of the map:

$$\sigma_\rho^2 = \frac{1}{V^2} \sum_{\mathbf{h}\neq 000} |F(\mathbf{h})|^2. \qquad (15.3.5.1)$$

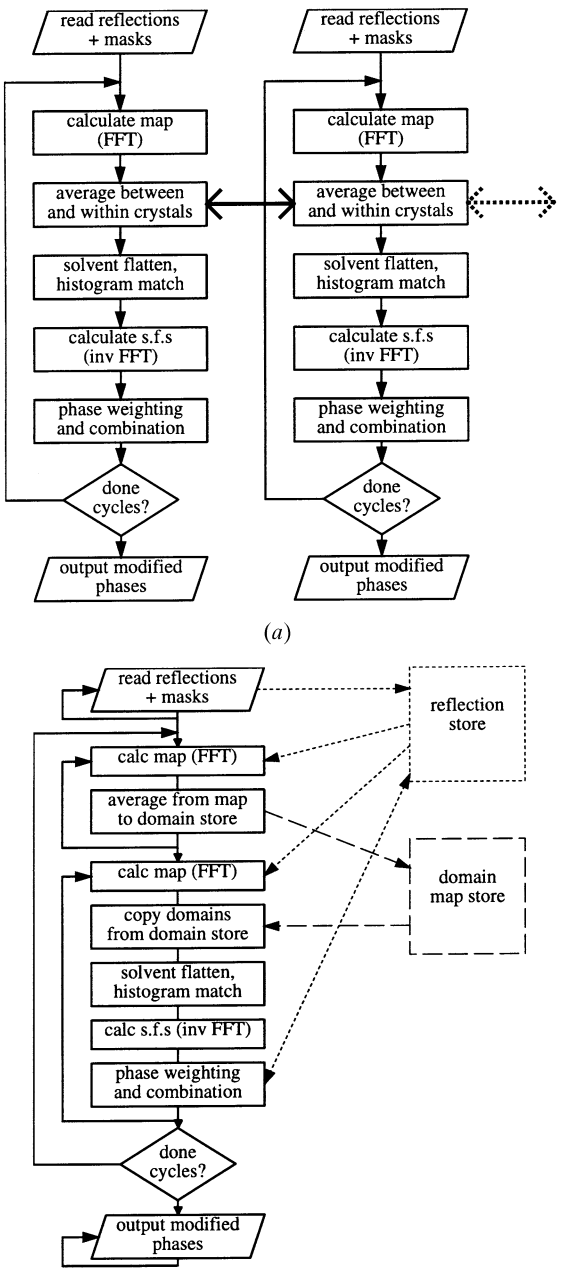
Thus, the sum of the intensities in a particular resolution shell is proportional to the difference in variance of maps calculated with and without that shell of data. The empirical curve is therefore calculated from the variance in the protein regions of a group of known structures, calculated as a function of resolution. The curve is scaled to the protein volume of the current structure, and a correction is made for the solvent, which is assumed to be flat.

The overall temperature factor is removed, and an absolute scale is imposed by fitting the data to this curve. The use of sharpened F's (with no overall temperature factor) is necessary


read
Fobs,σF,ϕ,w
read/calculate averaging mask
scale data to absolute, B=0
chose starting set reflections; set cycle n=1
read/calculate solvent mask
Calculate map ρ=FFT[wnF exp(iϕn)]
Modify map: ρ–>ρmod average molecules flatten solvent histogram match protein
Calculate sf's Fmod exp(iϕmod)=FFT[ρmod]
Calculate weight wmod from agreement of Fobs, Fmod
Combine phases/weights: P[ϕn+1]=P[ϕn]P[ϕmod]
n=n+1 n<ncycles?
Introduce new reflections
write ϕn,wn (+masks)
read
Fobs,σF,ϕ,w
read/calculate averaging mask
scale data to absolute, B=0
chose starting set reflections; set cycle n=1
read/calculate solvent mask
set m=0
Set F(h)=0 for all h in Sm
Calculate map ρ=FFT[wnF exp(iϕn)]
Modify map: ρ–>ρmod average molecules flatten solvent histogram match protein
Calculate sf's Fmod exp(iϕmod)=FFT[ρmod]
Fomit(h)=Fmod(h) h in Sm
m=m+1 m<Nsets?
Calculate weight womit from agreement of Fobs, Fomit
Combine phases/weights: P[ϕn+1]=P[ϕn]P[ϕomit]
n=n+1 n<Ncycles?
write ϕn,wn (+masks)

(*a*) (*b*)

Figure 15.3.5.1
(*a*) Flow chart for a simple *DM* calculation with free-Sim phase combination. (*b*) Flow chart for a simple *DM* calculation with reflection-omit phase combination.


read reflections + masks
calculate map (FFT)
average between and within crystals
solvent flatten, histogram match
calculate s.f.s (inv FFT)
phase weighting and combination
done cycles?
output modified phases
read reflections + masks
reflection store
calc map (FFT)
average from map to domain store
domain map store
calc map (FFT)
copy domains from domain store
solvent flatten, histogram match
calc s.f.s (inv FFT)
phase weighting and combination
done cycles?
output modified phases

(*a*)

(*b*)

Figure 15.3.5.2

(*a*) Conceptual flowchart for a *DMMULTI* multi-crystal calculation. (*b*) Actual flow chart for a *DMMULTI* multi-crystal calculation.

for histogram matching and often increases the power of averaging for phase extension.

Since the solvent content is used in scaling the data, it is important that this value be entered correctly. However, the volume of the solvent mask may be varied independently of the true solvent content, as discussed in Section 15.3.3.

15.3.5.2. *Solvent-mask determination*

If the user does not supply a solvent mask, the solvent mask is calculated by Wang's (1985) method, using the reciprocal-space approach of Leslie (1987). A number of variants on this algorithm are implemented; however, the parameter that affects the quality of the solvent mask most dramatically is the radius of the smoothing function (Chapter 15.1). This parameter may be estimated empirically by

$$r_{\rm Wang} = 2r_{\max}\overline{w}^{1/3}, \qquad (15.3.5.2)$$

where $r_{\max}$ is the resolution limit of the observed amplitudes, and $\overline{w}$ is the mean figure of merit over the same reflections (with $w = 0$ for unphased reflections).

Once the smoothed map has been determined, cutoff values are chosen to divide the map into protein and solvent regions. If the protein boundary is poorly defined, the user may specify protein, solvent and excluded volumes, in which case two cutoffs are specified and the intermediate region is marked as neither protein nor solvent.

15.3.5.3. *Averaging-mask determination*

If the user does not supply an averaging mask, it is determined by a local correlation method (Vellieux *et al.*, 1995). A large region covering 27 unit cells is selected, and the local correlation between the maps before and after rotation by one of the noncrystallographic symmetry operators is calculated. The largest contiguous region that is in agreement among different NCS operators is isolated from the local correlation map, and a finer local correlation map is calculated over this volume. This process is iterated until a good mask with a detailed boundary is found.

This approach is fully automatic, except in the case where a noncrystallographic symmetry operator intersects a crystallographic symmetry operator, in which case the mask is not uniquely defined, and some user intervention may be required. The method is robust, and by increasing the radius of the sphere within which the local correlation is calculated, it may be used with very poor maps (Cowtan & Main, 1998). The method is easily extended to include information from multiple averaging operators.

15.3.5.4. *Fourier transforms*

For simplicity of coding, all Fourier transforms are performed in core using real-to-Hermitian and Hermitian-to-real fast Fourier transforms (FFTs). The data are expanded to space group *P*1 before calculating a map and averaged back to a reciprocal asymmetric unit after inverse transformation. Most of the map modifications preserve crystallographic symmetry, so restricted phases are not constrained except during phase combination.

15.3.5.5. *Histogram matching*

The target histograms are calculated from the protein regions of several stationary-atom structures at resolutions from 6 to 1.5 Å, according to the method described by Zhang & Main (1990*a*). The histogram variances should be consistent with the map variances used in scaling the data. The resolution of the target histogram can be accurately matched to the data resolution by averaging the target histograms on either side of the current resolution.

15.3.5.6. *Averaging*

Averaging is performed using a single-step approach (Rossmann *et al.*, 1992), in which every copy of the molecule in a 'virtual' asymmetric unit is averaged with every other copy. Density values are obtained at non-grid positions using a 27-point quadratic spectral spline interpolation. A sharpened map is first calculated by dividing by the Fourier transform of the quadratic spline function. The same spline function is then convoluted with the sharpened map to obtain the density value at an arbitrary coordinate (Cowtan & Main, 1998). This approach gives very accurate interpolation from a coarse grid map with relatively little computation and additionally provides gradient information for the refinement of averaging operators.

15.3.5.7. Multi-crystal averaging

The multi-crystal averaging calculation in *DMMULTI* is equivalent to several single-crystal averaging calculations running simultaneously, with the exception that during the averaging step, the molecule density is averaged across every copy in every crystal form. This average is weighted by the mean figure of merit of each crystal form; this allows the inclusion of unphased crystal forms, since in the first cycle they will have zero weight and therefore not disrupt the phasing that is already present. In subsequent cycles, the unphased form contains phase information from the back-transformed density.

This technique can be extremely useful, since adding a new crystal form usually provides considerably more phase information than adding a new derivative if the cross-rotation and translation functions can be solved.

In the multi-crystal case, averaging is performed using a two-step approach, first building an averaged molecule from all the copies in all crystal forms, then replacing the density in each crystal form with the averaged values. This approach is computationally more efficient when there are many copies of the molecule.

The conceptual flow chart of simultaneous density-modification calculations across multiple crystal forms is shown in Fig. 15.3.5.2(*a*); in practice, this scheme is implemented using a single process and looping over every crystal form at each stage (Fig. 15.3.5.2*b*). Maps are reconstructed from a large data object containing all the reflection data in every crystal form. Averaging is performed using a second data object containing maps of each averaging domain. By this means, an arbitrary number of domains may be averaged across an arbitrary number of crystal forms.

Multi-crystal averaging has been particularly successful in solving structures from very weak initial phasing, since the data redundancy is usually higher than for single-crystal problems.

KDC acknowledges the support of the UK BBSRC (grant No. 87/B03785). KYJZ acknowledges the National Institutes of Health for grant support (GM55663).

References

Abrahams, J. P. (1997). *Bias reduction in phase refinement by modified interference functions: introducing the γ correction. Acta Cryst.* D**53**, 371–376.

Abrahams, J. P. & Leslie, A. G. W. (1996). *Methods used in the structure determination of bovine mitochondrial F_2 ATPase. Acta Cryst.* D**52**, 30–42.

Brünger, A. T. (1992). *X-PLOR.* Version 3.1. *A system for X-ray crystallography and NMR.* Yale University Press, New Haven.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Ab initio phase determination for spherical viruses: parameter determination for spherical-shell models. Acta Cryst.* A**48**, 301–312.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

Cowtan, K. D. & Main, P. (1996). *Phase combination and cross validation in iterated density-modification calculations. Acta Cryst.* D**52**, 43–48.

Cowtan, K. D. & Main, P. (1998). *Miscellaneous algorithms for density modification. Acta Cryst.* D**54**, 487–493.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Kabsch, W. (1976). *A solution for the best rotation to relate two sets of vectors. Acta Cryst.* A**32**, 922–923.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography. Methods Enzymol.* **277**, 269–305.

Leslie, A. G. W. (1987). *A reciprocal-space method for calculating a molecular envelope using the algorithm of B. C. Wang. Acta Cryst.* A**43**, 134–136.

Matthews, B. W. (1968). *Solvent content of protein crystals. J. Mol. Biol.* **33**, 491–497.

Matthews, B. W. (1974). *Determination of molecular weight from protein crystals. J. Mol. Biol.* **82**, 513–526.

Navaza, J. (1994). *AMoRe: an automated package for molecular replacement. Acta Cryst.* A**50**, 157–163.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Rossmann, M. G., McKenna, R., Tong, L., Xia, D., Dai, J.-B., Wu, H., Choi, H.-K. & Lynch, R. E. (1992). *Molecular replacement real-space averaging. J. Appl. Cryst.* **25**, 166–180.

Schuller, D. J. (1996). *MAGICSQUASH: more versatile non-crystallographic averaging with multiple constraints. Acta Cryst.* D**52**, 425–434.

Tsao, J., Chapman, M. S. & Rossmann, M. G. (1992). *Ab initio phase determination for viruses with high symmetry: a feasibility study. Acta Cryst.* A**48**, 293–301.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification. J. Appl. Cryst.* **28**, 347–351.

Wang, B. C. (1985). *Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol.* **115**, 90–112.

Zhang, K. Y. J. & Main, P. (1990*a*). *Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. Acta Cryst.* A**46**, 41–46.

Zhang, K. Y. J. & Main, P. (1990*b*). *The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. Acta Cryst.* A**46**, 377–381.

International Tables for Crystallography (2012). Vol. F, Chapter 16.1, pp. 413–432.

PART 16. DIRECT METHODS

Chapter 16.1. *Ab initio* phasing

G. M. Sheldrick, C. J. Gilmore, H. A. Hauptman, C. M. Weeks, R. Miller and I. Usón

16.1.1. Introduction

Ab initio methods for solving the crystallographic phase problem rely on diffraction amplitudes alone and do not require prior knowledge of any atomic positions. General features that are not specific to the structure in question (*e.g.* the presence of α-helices, disulfide bridges or solvent regions) can, however, be utilized. For the last four decades, most small-molecule structures have been routinely solved by *direct methods*, a class of *ab initio* methods in which probabilistic phase relations are used to derive reflection phases from the measured amplitudes. The direct solution of new macromolecular structures in this way has, however, been limited to a few special cases involving relatively small macromolecules, unusually high-resolution data and, often, the presence of heavier atoms [which might also have been suitable for single-wavelength anomalous diffraction (SAD) or multiple-wavelength anomalous diffraction (MAD) phasing]. However, the same procedures can be applied at much lower resolution for the location of heavy-atom substructures, an essential step in the experimental phasing of macromolecules in the widely used SAD, single isomorphous replacement including anomalous scattering (SIRAS), multiple isomorphous replacement (MIR) and MAD methods. Indeed, substructure-based phasing now accounts for most direct-methods applications to macromolecules. Since three closely related dual-space direct methods computer programs (*SnB*, *SHELXD* and *HySS*) are currently used in the large majority of such applications, we will concentrate on this approach and then describe more briefly some other promising approaches, including one that does not require high-resolution data, a related molecule as search fragment or heavier atoms and should, therefore, be applicable to at least a quarter of the protein structures in the Protein Data Bank (PDB).

16.1.1.1. *Data resolution*

Direct methods of crystal structure determination have wholly transformed small-molecule crystallography in the past two decades. The same cannot be said for macromolecular crystallography, although there have been very significant advances in the area of substructure determination. The reasons for the success with small molecules are:

(1) automatic and easy to use software is readily and freely available [*e.g.* *SHELXS* and *SHELXD* (Sheldrick, 1990, 2008; Usón & Sheldrick, 1999), *SnB* (Miller *et al.*, 1994; Weeks & Miller, 1999*a*), *SIR2004* (Burla *et al.*, 2005), and *SUPERFLIP* (Palatinus & Chapuis, 2007)];

(2) the high quality and, in particular, high resolution of data now collected from both laboratory sources and synchrotron facilities; and

(3) data sets are complete with few missing reflections.

Why do data resolution and data quality matter? To understand this, we need to examine a rule proposed by Sheldrick (1990): *Experience with a large number of structures has led us to formulate the empirical rule that, if fewer than half the number of theoretically measurable reflections in the range 1.1 to 1.2 Å are 'observed' [i.e. have $F > 4\sigma(F)$], it is very unlikely that the structure can be solved by direct methods. This critical ratio may be reduced somewhat for centrosymmetric structures and structures containing heavy atoms.*

Morris and Bricogne (2003) offer valuable structural insights into this rule that are instructive for this chapter. By examining the averaged squared normalized structure-factor amplitudes of more than 700 high-resolution (<2.0 Å) structures as a function of data resolution, they found that there is always a pronounced maximum around 1.1 Å, a smaller one around 2.1 Å, and a further pronounced one at ~4.5 Å as shown in Fig. 16.1.1.1. The shape of these curves can be related back to a sinc function transformation which links the intensities of normalized structure-factor profiles and the radial pair distribution function. The peak at 1.1 Å can be shown to arise from bonded distances of ~1.5 Å and non-bonded distances of ~2.4 Å; every protein can be shown to contain distance beats of 1.1 Å arising from these. The net result is a systematic reduction in the expectation value of $|E|^2$ to about 1.25 Å and only then does it rise again. At 1.1–1.2 Å, the resolution is sufficient to reproduce a radial distance distribution with suitably separated peaks, and this gives not only atomicity, but also the stereochemical regularities necessary for the successful application of direct methods. To exacerbate matters further, the fundamental equations of direct methods have variances with a $1/N^{1/2}$ dependence, where N is the number of atoms in the unit cell. Morris and Bricogne make the matter clear: direct methods in their current formulation will always struggle with macromolecular data. This said, however, there are two significant and general uses of direct methods in macromolecular crystallography:

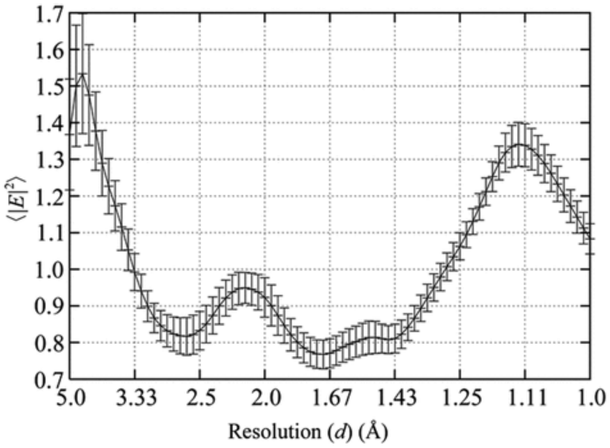
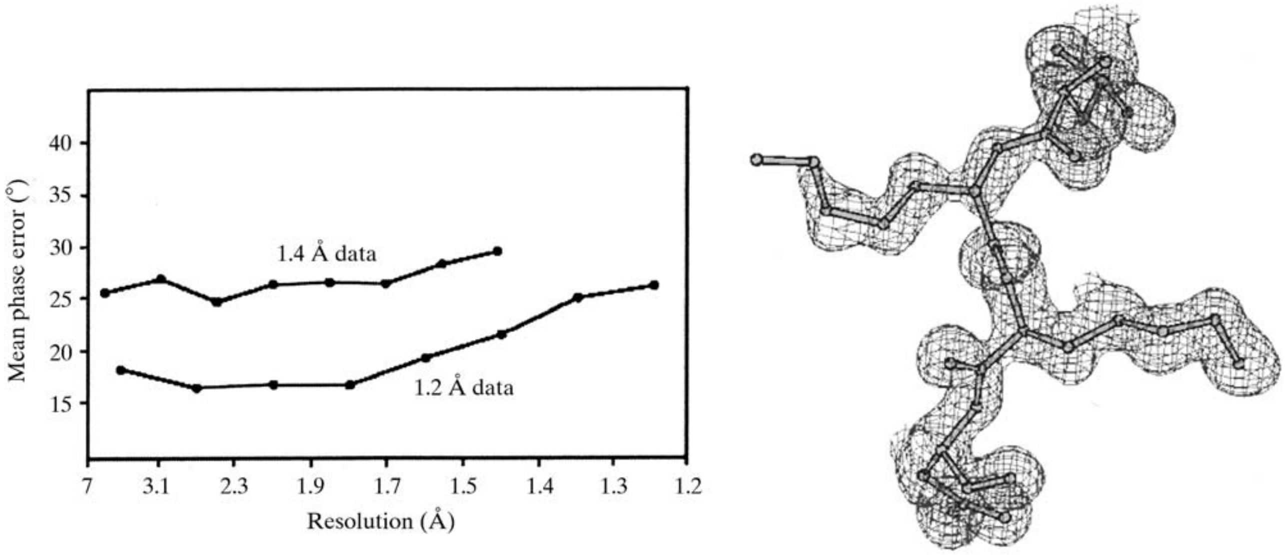


Figure 16.1.1.1
Averaged squared normalized structure-factor amplitudes over 700 protein structures with standard deviations calculated from the population of individual $|E|^2$ profiles (from Morris & Bricogne, 2003).

Copyright © 2012 International Union of Crystallography



(*a*) (*b*)

Figure 16.1.1.2
(*a*) Mean phase error as a function of resolution for the two independent *ab initio SHELXD* solutions of the previously unsolved protein hirustasin. Either the 1.2 Å or the 1.4 Å native data set led to solution of the structure. (*b*) Part of the hirustasin molecule from the 1.4 Å room-temperature data after one round of *B*-value refinement with fixed coordinates.

(1) *Ab initio* structure solution with atomic resolution data: the whole crystal structure solution is required with most of the atomic sites sufficiently defined for least-squares refinement. Sheldrick's rule applies rigorously here with very few exceptions in the literature. There have been isolated successes at lower resolutions, but these mostly involved the presence of heavier atoms or data with truncated resolution from crystals that would have diffracted further.

(2) Substructure solution: the determination of the positions of the heavy atoms only (often Se from selenomethionine, but also quite frequently heavy-atom salts, complexes and clusters). Sheldrick's rule is substantially relaxed and structure solutions at data resolutions of 5–6 Å are possible. This follows from the arguments of Morris and Bricogne: the distance beats of 1.1 Å are irrelevant to the substructure. By focusing on the heavy atoms, the complexity of the structure is much reduced, the distance between atoms is larger and the solution of the phase problem becomes easier.

The importance of the presence of several atoms heavier than oxygen for increasing the chance of obtaining a solution by the program *SnB* at resolutions less than 1.2 Å was noticed for truncated data from vancomycin and the 289-atom structure of conotoxin EpI (Weeks & Miller, 1999*b*). The results of *SHELXD* application to hirustasin, which contains ten sulfur and 457 carbon, nitrogen and oxygen atoms in the asymmetric unit, are consistent with this (Usón *et al.*, 1999). The 55-amino-acid protein hirustasin could be solved by *SHELXD* using either 1.2 Å low-temperature data or 1.4 Å room-temperature data. However, as shown in Fig. 16.1.1.2(*a*), the mean phase error (MPE) is significantly better for the 1.2 Å data over the whole resolution range. Although small-molecule interpretation based on peak positions worked well for the 1.2 Å solution (overall MPE $= 18°$), standard protein chain tracing was required for the 1.4 Å solution (overall MPE $= 26°$). As is clear from the corresponding electron-density map (Fig. 16.1.1.2*b*), *SHELXD* produced easily interpreted protein density even when bonded atoms are barely resolved from each other.

16.1.1.2. Data completeness

The relative effects of accuracy, completeness and resolution on *Shake-and-Bake* success rates using *SnB* for three large *P*1 structures were studied by computing error-free data using the known atomic coordinates (Xu *et al.*, 2000). The results of these studies, presented in Table 16.1.1.1, show that experimental error contributed nothing of consequence to the low success rates for vancomycin and lysozyme. However, completing the vancomycin data up to the maximum measured resolution of 0.97 Å resulted in a substantial increase in success rate which was further improved to an astounding success rate of 80% when the data were expanded to 0.85 Å. As a result of problems with overloaded reflections, the experimental vancomycin data did not include any data at 10 Å resolution or lower. A total of 4000 reflections were phased in the process of solving this structure with the experimental data. Some of these data were then replaced with the largest error-free magnitudes chosen from the missing reflections at several different resolution limits. The results in Table 16.1.1.2 show a tenfold increase in success rate when only 200 of the largest missing magnitudes were supplied, and it made no difference whether these reflections had a maximum resolution of 2.8 Å or were chosen randomly from the whole 0.97 Å sphere. The moral of this story is that, *when collecting synchrotron data for direct methods, it pays to take a second pass using a shorter exposure time to fill in the low-resolution data*.

16.1.1.3. Summary

The basic theory underlying direct methods has been summarized in an excellent chapter (Giacovazzo, 2008) in *International Tables for Crystallography* Volume B (Chapter 2.2) to which the reader is referred for details. Suffice it for this chapter to say that classical direct methods attempt to reconstruct the missing phase information using native data alone by utilizing direct relationships between the crystallographic phases without any *a priori* structural information.

From a historical perspective, the first successful applications of direct methods to native data for structures that could legiti-

Table 16.1.1.1
Success rates for three $P1$ structures illustrate the importance of using complete data to the highest possible resolution

	Vancomycin	Alpha-1	Lysozyme
Atoms	547	471	~1200
Completeness (%)	80.2	85.6	68.3
Resolution (Å)	0.97	0.90	0.85
Parameter shift	112.5°, 1	90°, 2	90°, 2
Success rates (%)			
Experimental	0.25	14	0
Error-free	0.2	19	0
Error-free complete	14	29	0.8
Error-free complete extended to 0.85 Å	80	42	—

References: vancomycin: Loll *et al.* (1998); alpha-1: Privé *et al.* (1999); lysozyme: Deacon *et al.* (1998).

Table 16.1.1.2
Improving success rates by 'completing' the vancomycin data

Error-free reflections added	Success rate (%)
0	0.25
100 (3.5 Å)	0.3
200 (2.8 Å)	2.1
200 (0.97 Å)	2.4
400 (1.3 Å)	8.2
800 (1.1 Å)	11.1

mately be regarded as small macromolecules came from the *Shake-and-Bake* method and the associated *SnB* software (Weeks *et al.*, 1993). The distinctive feature of this procedure is the repeated and unconditional alternation of reciprocal-space phase refinement ('shaking') with a complementary real-space process that seeks to improve phases by applying constraints ('baking'). The first previously unknown structures determined by *Shake-and-Bake* were two forms of the 100-atom peptide ternatin (Miller *et al.*, 1993) and, so far, the largest previously unsolved structure solved by direct methods with no atom heavier than oxygen is probably feglymycin, with 1026 unique non-hydrogen atoms and data to 1.10 Å resolution (Bunkóczi *et al.*, 2005).

Using direct methods and accurately measured data, it is now possible to solve heavy-atom substructures of well over 100 atoms. For a state-of-the-art example, see von Delft *et al.* (2003), where a substructure of 160 Se atoms was solved in the product-bound *E. coli* KPHMT using *SnB*. A total of 120 sites were correctly located, allowing the remainder to be located by *SHARP* (de La Fortelle & Bricogne, 1997); in later tests, *SHELXD* was able to find 152 of the sites. For a review of the phase problem in the context of other developments, the reader is referred to a general overview by Dauter (2006).

The present chapter focuses on those aspects of direct methods that have proven useful for larger molecules (more than 250 independent non-H atoms) or are unique to the macromolecular field. These include direct-methods applications that utilize anomalous-dispersion measurements or multiple diffraction patterns [*i.e.* single isomorphous replacement (SIR), SAD and MAD] to locate substructures at resolutions typically in the range 2.0–3.5 Å, although lower-resolution data are sometimes adequate. A formal integration of the probabilistic machinery of direct methods with isomorphous replacement and anomalous dispersion was initiated in 1982 (Hauptman, 1982*a*,*b*). Although practical applications of this and subsequent related theory have been limited so far, this approach might prove relevant in the future. Similarly, the combination of direct methods with multiple-beam diffraction might also play a role (Weckert *et al.*, 1993).

Table 16.1.2.1
Theoretical values pertaining to $|E|$'s

	Centrosymmetric	Noncentrosymmetric
Average $\lvert E\rvert^2$	1.000	1.000
Average $\lvert \lvert E^2\rvert - 1\rvert$	0.968	0.736
Average $\lvert E\rvert$	0.798	0.886
$\lvert E\rvert > 1$ (%)	32.0	36.8
$\lvert E\rvert > 2$ (%)	5.0	1.8
$\lvert E\rvert > 3$ (%)	0.3	0.01

16.1.2. Normalized structure-factor magnitudes

For purposes of direct-methods computations, the usual structure factors, $F_{\mathbf{H}}$, are replaced by the *normalized structure factors* (Hauptman & Karle, 1953),

$$E_{\mathbf{H}} = |E_{\mathbf{H}}| \exp(i\varphi_{\mathbf{H}}),$$

$$|E_{\mathbf{H}}| = \frac{|F_{\mathbf{H}}|}{\langle |F_{\mathbf{H}}|^2 \rangle^{1/2}} = \frac{k\langle \exp[-B_{\rm iso}(\sin\theta)^2/\lambda^2] \rangle^{-1} |F_{\mathbf{H}}|_{\rm meas}}{\left(\varepsilon_{\mathbf{H}} \sum_{j=1}^{N} f_j^2\right)^{1/2}}, \qquad (16.1.2.1)$$

where the angle brackets indicate probabilistic or statistical expectation values, the $|E_{\mathbf{H}}|$ and $|F_{\mathbf{H}}|$ are structure-factor magnitudes, the $\varphi_{\mathbf{H}}$ are the corresponding phases, k is the absolute scaling factor for the measured magnitudes, $B_{\rm iso}$ is an overall isotropic atomic mean-square displacement parameter, the f_j are the atomic scattering factors for the N atoms in the unit cell, and the $\varepsilon_{\mathbf{H}} \geq 1$ are factors that account for multiple enhancement of the average intensities for certain special reflection classes due to space-group symmetry (Shmueli & Wilson, 2008). The condition $\langle |E|^2 \rangle = 1$ is always imposed. Unlike $\langle |F_{\mathbf{H}}| \rangle$, which decreases as $\sin(\theta)/\lambda$ increases, the values of $\langle |E_{\mathbf{H}}| \rangle$ are constant for concentric resolution shells. Thus, the normalization process places all reflections on a common basis, and this is a great advantage with regard to the probability distributions that form the foundation for direct methods. Normalizing a set of reflections by means of equation (16.1.2.1) does not require any information about atomic positions. However, if some structural information, such as the configuration, orientation, or position of certain atomic groupings, is available, then this information can be applied to obtain a better model for the expected intensity distribution (Main, 1976). The distribution of values is, in principle and often in practice, independent of the unit-cell size and contents, but it does depend on whether a centre of symmetry is present, as shown in Table 16.1.2.1.

Direct-methods applications having the objective of locating SIR or SAD substructures require the computation of normalized *difference* structure-factor magnitudes, $|E_\Delta|$. This can, for example, be accomplished with the following series of programs from Blessing's data-reduction and error-analysis routines (*DREAR*): *LEVY* and *EVAL* for structure-factor normalization as specified by equation (16.1.2.1) (Blessing *et al.*, 1996), *LOCSCL* for local scaling of the SIR and SAD magnitudes (Matthews & Czerwinski, 1975; Blessing, 1997), and *DIFFE* for computing the actual difference magnitudes (Blessing & Smith, 1999). The *SnB* program (see Section 16.1.12.4) provides a convenient interface to the *DREAR* suite.

16.1.2.1. SIR differences

Given the individual normalized structure-factor magnitudes for native structures ($|E_{\rm nat}|$) and for structures containing heavy atoms ($|E_{\rm der}|$), as well as the atomic scattering factors $|f_j| = |f_j^0 + f_j' + if_j''| = [(f_j^0 + f_j')^2 + (f_j'')^2]^{1/2}$ which allow for the possibility of anomalous scattering, then greatest-lower-bound estimates of SIR difference-E magnitudes are

$$|E_\Delta| = \frac{\left|\left(\sum_{j=1}^{N_{\rm der}} |f_j|^2\right)^{1/2} |E_{\rm der}| - \left(\sum_{j=1}^{N_{\rm nat}} |f_j|^2\right)^{1/2} |E_{\rm nat}|\right|}{q\left[\left(\sum_{j=1}^{N_{\rm der}} |f_j|^2\right) - \left(\sum_{j=1}^{N_{\rm nat}} |f_j|^2\right)\right]^{1/2}}, \quad (16.1.2.2)$$

where $q = q_0 \exp(q_1 s^2 + q_2 s^4)$ is a least-squares-fitted empirical renormalization scaling function, dependent on $\sin(\theta)/\lambda$, that imposes the condition $\langle |E_\Delta|^2 \rangle = 1$ and serves to define q_0, q_1 and q_2.

16.1.2.2. SAD differences

Given Friedel pairs of normalized structure-factor magnitudes ($|E_{+\mathbf{H}}|$, $|E_{-\mathbf{H}}|$) and the atomic scattering factors (f^0, f_j' and f_j''), then the greatest-lower-bound estimates of SAD difference $|E|$'s are

$$|E_\Delta| = \frac{\left[\sum_{j=1}^{N} (f_j^0 + f_j')^2 + (f_j'')^2\right]^{1/2} \big||E_{+\mathbf{H}}| - |E_{-\mathbf{H}}|\big|}{2q\left[\sum_{j=1}^{N} (f_j'')^2\right]^{1/2}}, \quad (16.1.2.3)$$

where, again, q is an empirical renormalization scaling function that imposes the condition $\langle |E_\Delta|^2 \rangle = 1$.

In the case of MAD data, the anomalous differences for all wavelengths can be combined in the form of F_A structure factors as described in Chapter 14.2 of this volume by Smith & Hendrickson. In the *SHELX* system, the program *SHELXC* prepares the $|F_A|$ values and the phase shifts α for SAD, SIR, SIRAS, RIP (radiation-damage-induced phasing) and MAD experimental phasing (Sheldrick, 2008, 2010), and the corresponding $|E_A|$ values are derived by *SHELXD*.

16.1.2.3. Difference intensities and direct methods

There are, of course, difficulties with normalized difference intensities. Direct methods normalize the data to give difference E magnitudes, and the macromolecular temperature factors incorporated in this process can be a source of error. The temperature-factor correction, often large for macromolecules, can take small, statistically dubious, amplitude differences at high resolution when working with anomalous difference data and magnify them into large normalized structure-factor differences that are then used in direct methods. These methods are very sensitive to such errors and will often fail unless a suitable resolution truncation limit is imposed. Usually this is around 0.5 Å larger than the experimental diffraction limit of the data or up to the resolution shell with $I/\sigma(I)$ = 10. Morris *et al.* (2004) have extended their work on profiles to devise a method of normalization based on $\langle |E|^2 \rangle$ profiles, and this should be more robust than traditional methods.

16.1.3. Starting the phasing process

The phase problem of X-ray crystallography may be defined as the problem of determining the phases φ of the normalized structure factors E when only the magnitudes $|E|$ are given. Owing to the atomicity of crystal structures and the redundancy of the known magnitudes, the phase problem is overdetermined and is, therefore, solvable in principle. This overdetermination implies the existence of relationships among the E's and, since the magnitudes $|E|$ are presumed to be known, there exist identities among the phases that are dependent on the known magnitudes alone. The techniques of probability theory lead to the joint probability distributions of arbitrary collections of E from which the conditional probability distributions of selected sets of phases, given the values of suitably chosen magnitudes $|E|$, may be inferred.

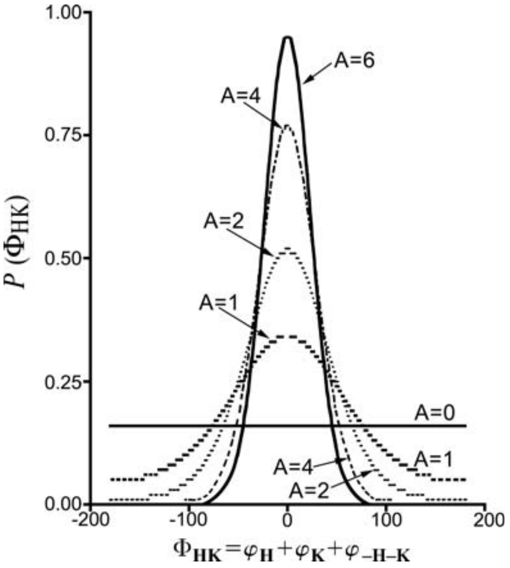


Figure 16.1.3.1

The conditional probability distribution, $P(\Phi_{\mathbf{HK}})$, of the three-phase structure invariants, $\Phi_{\mathbf{HK}}$, having associated parameters $A_{\mathbf{HK}}$ with values of 0, 1, 2, 4 and 6. When A $\simeq$ 0, all values of $\Phi_{\mathbf{HK}}$ are equally likely, and no information useful for phase determination is available. However, the sum of the three phases for most invariants with A $\simeq$ 6 is close to 0°, and an estimate of one phase can be made if the other two are known.

16.1.3.1. Structure invariants

The magnitude-dependent entities that constitute the foundation of direct methods are linear combinations of phases called *structure invariants*. The term 'structure invariant' stems from the fact that the values of these quantities are independent of the choice of origin. The most useful of the structure invariants are the three-phase or *triplet invariants*,

$$\Phi_{\mathbf{HK}} = \varphi_{\mathbf{H}} + \varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}}, \quad (16.1.3.1)$$

the conditional probability distribution (Cochran, 1955), given $A_{\mathbf{HK}}$, of which is

$$P(\Phi_{\mathbf{HK}}) = [2\pi I_0(A_{\mathbf{HK}})]^{-1} \exp(A_{\mathbf{HK}} \cos \Phi_{\mathbf{HK}}), \quad (16.1.3.2)$$

where

$$A_{\mathbf{HK}} = (2/N^{1/2})|E_{\mathbf{H}} E_{\mathbf{K}} E_{\mathbf{H}+\mathbf{K}}| \quad (16.1.3.3)$$

and N is the number of atoms, here presumed to be identical, in the primitive unit cell. This distribution is illustrated in Fig. 16.1.3.1. The expected value of the cosine of a particular triplet, $\Phi_{\mathbf{HK}}$, is given by the ratio of modified Bessel functions, $I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})$.

Estimates of the invariant values are most reliable when the normalized structure-factor magnitudes ($|E_{\mathbf{H}}|$, $|E_{\mathbf{K}}|$ and $|E_{-\mathbf{H}-\mathbf{K}}|$) are large and the number of atoms in the corresponding primitive unit cell, N, is small. This is one important reason why direct phasing is more difficult for macromolecules than it is for small molecules. Four-phase or quartet invariants have proven helpful in small-molecule structure determination, particularly when used passively as the basis for a figure of merit (DeTitta *et al.*, 1975). However, the reliability of these invariants, as given by their conditional probability distribution (Hauptman, 1975), is proportional to $1/N$, and they have not as yet been shown to be useful for macromolecular phasing. The reliability of higher-order invariants decreases even more rapidly as structure size increases.

16.1.3.2. *'Multisolution' methods and trial structures*

Successful crystal structure determination requires that sufficient phases be found such that a Fourier map computed using the corresponding structure factors will reveal the atomic positions. It is particularly important that the biggest terms (*i.e.*, largest $|E|$) be included in the Fourier series. Thus, the first step in the phasing process is to sort the reflections in decreasing order according to their $|E|$ values and to choose the number of large $|E|$ reflections that are to be phased. The second step is to generate the possible invariants involving these intense reflections and then to sort them in decreasing order according to their $A_{\mathbf{HK}}$ values. Those invariants with the largest $A_{\mathbf{HK}}$ values are retained in sufficient number to achieve the desired overdetermination. *Ab initio* phase determination by direct methods requires not only a set of invariants, the average values of the cosines of which are presumed to be known, but also a set of starting phases. Therefore, the third step in the phasing process is the assignment of initial phase values. If enough pairs of phases, $\varphi_{\mathbf{K}}$ and $\varphi_{-\mathbf{H}-\mathbf{K}}$, are known, the structure invariants can then be used to generate further phases ($\varphi_{\mathbf{H}}$) which, in turn, can be used to evaluate still more phases. Repeated iterations will permit most reflections with large $|E_{\mathbf{H}}|$ to be phased.

Depending on the space group, a small number of phases can be assigned arbitrarily in order to fix the origin position and, in noncentrosymmetric space groups, the enantiomorph or polar-axis direction. However, except for the simplest structures, these reflections provide an inadequate foundation for further phase development. Historically, a 'multisolution' or multi-trial approach (Germain & Woolfson, 1968) was taken in which other reflections are each assigned many different starting values in the hope that one or more of the resultant phase combinations will lead to a solution. Solutions, if they occurred, were identified on the basis of some suitable figure of merit. Although phases can be evaluated sequentially, the order determined by a so-called convergence map (Germain *et al.*, 1970), it has become standard in recent years to use a random-number generator to assign initial values to all available phases from the outset (Baggio *et al.*, 1978; Yao, 1981). A variant of this procedure is to use the random-number generator to assign initial coordinates to the atoms in the trial structures and then to obtain initial phases from a structure-factor calculation. In addition, some dual-space programs [*SHELXD* (Schneider & Sheldrick, 2002), *HySS* (Grosse-Kunstleve & Adams, 2003)] can also use the Patterson function to generate starting atoms or phases (see Section 16.1.8).

16.1.4. Reciprocal-space phase refinement or expansion (*shaking*)

Once a set of initial phases has been chosen, it must be refined against the set of structure invariants whose values are presumed known. In theory, any of a variety of optimization methods could be used to extract phase information in this way. However, so far only three purely reciprocal-space methods have been shown to be of practical value: tangent refinement, parameter-shift optimization of the minimal function, and maximization of Karle–Hauptman determinants (Section 16.1.12.7).

16.1.4.1. *The tangent formula*

The *tangent formula*,

$$\tan(\varphi_{\mathbf{H}}) = \frac{-\sum_{\mathbf{K}} |E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}| \sin(\varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}})}{\sum_{\mathbf{K}} |E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}| \cos(\varphi_{\mathbf{K}} + \varphi_{-\mathbf{H}-\mathbf{K}})}, \qquad (16.1.4.1)$$

(Karle & Hauptman, 1956), is the relationship used in conventional direct-methods programs to compute $\varphi_{\mathbf{H}}$ given a sufficient number of pairs $(\varphi_{\mathbf{K}}, \varphi_{-\mathbf{H}-\mathbf{K}})$ of known phases. It can also be used within the phase-refinement portion of the dual-space *Shake-and-Bake* procedure (Weeks, Hauptman *et al.*, 1994; Usón & Sheldrick, 1999). The variance associated with $\varphi_{\mathbf{H}}$ depends on $\sum_{\mathbf{K}} E_{\mathbf{H}}E_{\mathbf{K}}E_{-\mathbf{H}-\mathbf{K}}/N^{1/2}$ and, in practice, the estimate is only reliable for $|E_{\mathbf{H}}| \gg 1$ and for structures with a limited number of atoms (N). If equation (16.1.4.1) is used to redetermine previously known phases, the phasing process is referred to as *tangent-formula refinement*; if only new phases are determined, the phasing process is *tangent expansion*.

The tangent formula can be derived using the assumption of equal resolved atoms. Nevertheless, it suffers from the disadvantage that, in space groups without translational symmetry, it is perfectly fulfilled by a false solution with all phases equal to zero, thereby giving rise to the so-called 'uranium-atom' solution with one dominant peak in the corresponding Fourier synthesis. In conventional direct-methods programs, the tangent formula is often modified in various ways to include (explicitly or implicitly) information from the so-called 'negative' quartet invariants (Schenk, 1974; Hauptman, 1974; Giacovazzo, 1976) that are dependent on the smallest as well as the largest E magnitudes. Such modified tangent formulas do indeed largely overcome the problem of pseudosymmetric solutions for small N, but because of the dependence of quartet-term probabilities on $1/N$, they are little more effective than the normal tangent formula for large N.

16.1.4.2. *The minimal function*

Constrained minimization of an objective function like the *minimal function*,

$$R(\Phi) = \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}}\big\{\cos\Phi_{\mathbf{HK}} - [I_1(A_{\mathbf{HK}})/I_0(A_{\mathbf{HK}})]\big\}^2 \Big/ \sum_{\mathbf{H},\mathbf{K}} A_{\mathbf{HK}} \qquad (16.1.4.2)$$

(Debaerdemaeker & Woolfson, 1983; Hauptman, 1991; DeTitta *et al.*, 1994), provides an alternative approach to phase refinement or phase expansion. $R(\Phi)$ is a measure of the mean-square difference between the values of the triplets calculated using a particular set of phases and the expected values of the same triplets as given by the ratio of modified Bessel functions. The minimal function is expected to have a constrained global minimum when the phases are equal to their correct values for some choice of origin and enantiomorph (the minimal principle). Experimentation has thus far confirmed that, when the minimal function is used actively in the phasing process and solutions are produced, the final trial structure corresponding to the smallest value of $R(\Phi)$ is a solution provided that $R(\Phi)$ is calculated directly from the atomic positions before the phase-refinement step (Weeks, DeTitta *et al.*, 1994). Therefore, $R(\Phi)$ is also an extremely useful figure of merit. The minimal function can also include contributions from higher-order (*e.g.* quartet) invariants, although their use is not as imperative as with the tangent formula because the minimal function does not have a minimum when all phases are zero. In practice, quartets are rarely used in the minimal function because they increase the CPU time while adding little useful information for large structures.

The cosine function in equation (16.1.4.2) can also be replaced by other functions of the phases giving rise to alternative minimal functions. Examples include an exponential expression that has been found to give superior results for several $P1$ structures (Hauptman *et al.*, 1999). In addition, substructure determination using a very simple and computationally efficient modified minimal function,

$$m(\varphi) = 1 - (N_I/N_T) \tag{16.1.4.3}$$

(where I is an arbitrary interval $[-r, r]$, N_I is the number of triplets whose values lie in I and N_T is the total number of triplets), has been reported (Xu & Hauptman, 2004, 2006; Xu *et al.*, 2005) and incorporated into the *BnP* software (see Section 16.1.12.4).

16.1.4.3. Parameter shift

In principle, any minimization technique could be used to minimize $R(\Phi)$ by varying the phases. So far, a seemingly simple algorithm, known as parameter shift (Bhuiya & Stanley, 1963), has proven to be quite powerful and efficient as an optimization method when used within the *Shake-and-Bake* context to reduce the value of the minimal function. For example, a typical phase-refinement stage consists of three iterations or scans through the reflection list, with each phase being shifted a maximum of two times by 90° in either the positive or negative direction during each iteration. The refined value for each phase is selected, in turn, through a process which involves evaluating the minimal function using the original phase and each of its shifted values (Weeks, DeTitta *et al.*, 1994). The phase value that results in the lowest minimal-function value is chosen at each step. Refined phases are used immediately in the subsequent refinement of other phases. It should be noted that the parameter-shift routine is similar to that used in ψ-map refinement (White & Woolfson, 1975) and *XMY* (Debaerdemaeker & Woolfson, 1989).

16.1.5. Real-space constraints (*baking*)

For several decades, classical direct methods operated exclusively in reciprocal space, determining phases through statistical relationships between them. Only when this process had converged did the method move into real space by calculating one or more electron-density maps that were examined using stereochemical criteria. In macromolecular crystallography, density modification has always played a central role in phasing. A major advance in direct methods for macromolecules (and large molecules in general) occurred when density-modification methods were incorporated and adapted into the phasing procedure. They are often very simple: peaks which give rise to unrealistic geometries or which are too weak are removed, new structure factors are calculated and hence new phase angles are derived in an iterative process. (They can also be quite sophisticated as in *ACORN2*, which we will discuss in Section 16.1.12.1.) A consequence of this is that the once-clear dividing line between direct methods and other structure-solution techniques has become somewhat blurred.

Peak picking is a simple but powerful way of imposing an atomicity constraint. The potential for real-space phase improvement in the context of small-molecule direct methods was recognized by Karle (1968). He found that even a relatively small, chemically sensible, fragment extracted by manual interpretation of an electron-density map could be expanded into a complete solution by transformation back to reciprocal space and then performing additional iterations of phase refinement with the tangent formula. Automatic real-space electron-density-map interpretation in the *Shake-and-Bake* procedure consists of selecting an appropriate number of the largest peaks in each cycle to be used as an updated trial structure without regard to chemical constraints other than a minimum allowed distance between atoms. If markedly unequal atoms are present, appropriate numbers of peaks (atoms) can be weighted by the proper atomic numbers during transformation back to reciprocal space in a subsequent structure-factor calculation. Thus, *a priori* knowledge concerning the chemical composition of the crystal is utilized, but no knowledge of constitution is required or used during peak selection. It is useful to think of peak picking in this context as simply an extreme form of density modification appropriate when atomic resolution data are available. In theory, under appropriate conditions it should be possible within the dual-space direct-methods framework to replace peak picking by alternative density-modification procedures such as low-density elimination (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993) or solvent flattening (Wang, 1985). The imposition of physical constraints counteracts the tendency of phase refinement to propagate errors or produce overly consistent phase sets. Several variants of peak picking, which are discussed below, have been successfully employed within the framework of *Shake-and-Bake*.

16.1.5.1. Simple peak picking

In its simplest form, peak picking consists of simply selecting the top N_u E-map peaks where N_u is the number of unique non-H atoms in the asymmetric unit. This is adequate for true small-molecule structures. It has also been shown to work well for heavy-atom or anomalously scattering substructures where N_u is taken to be the number of expected substructure atoms (Smith *et al.*, 1998; Turner *et al.*, 1998). For larger structures ($N_u > 100$), it is likely to be better to select about $0.8N_u$ peaks, thereby taking into account the probable presence of some atoms that, owing to high thermal motion or disorder, will not be visible during the early stages of a structure determination. Furthermore, a study by Weeks & Miller (1999*b*) has shown that structures in the 250–1000-atom range which contain a half dozen or more moderately heavy atoms (*i.e.*, S, Cl, Fe) are more easily solved if only $0.4N_u$ peaks are selected. The only chemical information used at this stage is a minimum inter-peak distance, generally taken to be 1.0 Å. For substructure applications, a larger minimum distance (*e.g.* 3 Å) is more appropriate, provided that care is taken with disulfide bridges (Section 16.1.11).

16.1.5.2. Iterative peaklist optimization

An alternative approach to peak picking is to select approximately N_u peaks as potential atoms and then eliminate some of them, one by one, while maximizing a suitable figure of merit such as

$$P = \sum_{\mathbf{H}} |E_c^2|(|E_o^2| - 1). \tag{16.1.5.1}$$

The top N_u peaks are used as potential atoms to compute $|E_c|$. The atom that leaves the highest value of P is then eliminated. Typically, this procedure, which has been termed *iterative peaklist optimization* (Sheldrick & Gould, 1995), is repeated until only $2N_u/3$ atoms remain. Use of equation (16.1.5.1) may be regarded as a reciprocal-space method of maximizing the fit to the origin-removed sharpened Patterson function, and it has been used for this purpose in molecular replacement (Beurskens, 1981). Subject to various approximations, maximum-likelihood considerations also indicate that it is an appropriate function to maximize (Bricogne, 1998). Iterative peaklist optimization provides a higher percentage of solutions than simple peak picking, but it suffers from the disadvantage of requiring much more CPU time and so is less effective than the random-omit method described in the next section.

16.1.5.3. Random omit maps

A third peak-picking strategy involves selecting approximately $1.3N_u$ of the top peaks and eliminating some, but, in this case, the deleted peaks are chosen at random. Typically, one-third of the potential atoms are removed, and the remaining atoms are used to compute E_c. By analogy to the common practice in macromolecular crystallography of omitting part of a structure from a Fourier calculation in the hope of finding an improved position for the deleted fragment, this version of peak picking is described as *random omit*. This procedure helps to prevent the dual-space recycling from getting stuck in a local minimum and is thus an efficient search algorithm.

16.1.6. Fourier refinement

E-map recycling, but without phase refinement (Sheldrick, 1982, 1990; Kinneging & de Graaff, 1984), has been frequently used in conventional direct-methods programs to improve the completeness of the solutions after phase refinement. It is important to apply Fourier refinement to *Shake-and-Bake* solutions also because such processing significantly increases the number of resolved atoms, thereby making the job of map interpretation much easier. Since phase refinement *via* either the tangent formula or the minimal function requires relatively accurate invariants that can only be generated using the larger E magnitudes, a limited number of reflections are phased during the actual dual-space cycles. Working with a limited amount of data has the added advantage that less CPU time is required. However, if the current trial structure is the 'best' so far based on a figure of merit (either the minimal function or a real-space criterion), then it makes sense to subject this structure to Fourier refinement using additional data, thereby reducing series-termination errors. The correlation coefficient

$$\begin{aligned}\mathrm{CC} = {} & \left[\left(\sum wE_o^2E_c^2\sum w\right) - \left(\sum wE_o^2\sum wE_c^2\right)\right] \\ & \times \left\{\left[\left(\sum wE_o^4\sum w\right) - \left(\sum wE_o^2\right)^2\right]\right. \\ & \times \left.\left[\left(\sum wE_c^4\sum w\right) - \left(\sum wE_c^2\right)^2\right]\right\}^{-1/2} \qquad (16.1.6.1)\end{aligned}$$

(Fujinaga & Read, 1987), where weights $w = 1/[0.04 + \sigma^2(E_o)]$, has been found to be an especially effective figure of merit when used with all the data and is, therefore, suited for identifying the most promising trial structure at the end of Fourier refinement. Either simple peak picking or iterative peaklist optimization can be employed during the Fourier-refinement cycles in conjunction with weighted E maps (Sim, 1959). The final model can be further improved by isotropic displacement parameter ($B_{\rm iso}$) refinement for the individual atoms (Usón *et al.*, 1999) followed by calculation of the Sim (1959) or sigma-A (Read, 1986) weighted map. This is particularly useful when the requirement of atomic resolution is barely fulfilled, and it makes it easier to interpret the resulting maps by classical macromolecular methods.

16.1.7. Resolution enhancement: the 'free lunch' algorithm

Direct methods take a set of phases, refine them and also determine new ones. There is no reason, however, why they cannot be used to predict new amplitudes as well. If density modification of a real-space map is performed, then any process of real-space density modification will, following a Fourier transformation, give structure-factor amplitudes for reflections that were not used to generate it, and these can be outside the resolution limit. If direct methods require atomic resolution data, can we use these techniques to extrapolate structure factors (*i.e.* predict not only phases, but also amplitudes for missing data) and extend data resolution? The idea is not new, but it has been quite extensively studied in recent years. Sheldrick has termed such algorithms 'free lunch', with reference to the saying: 'There is no such thing as a free lunch'! In one example (Usón *et al.*, 2007), weak SIRAS starting phase information followed by density modification led to an $|F_o|$ weighted mean phase error (MPE) of 54° at 1.98 Å resolution, but when the density modification was performed with amplitude extrapolation to 1.0 Å, the MPE fell to 17°. Caliandro *et al.* (2005*a*,*b*) used Patterson or direct methods to obtain trial phases that are submitted to various density-modification methods. Following this, extrapolated phases were generated. This too transformed uninterpretable maps into a solution amenable to automatic tracing. Palatinus *et al.* (2007) used maximum entropy (ME) methods for amplitude extrapolation. In some ways these should be ideal for this purpose, and it is worth noting that ME maps have, *de facto*, optimal resolution enhancement built in, although they can be difficult to generate for large structures.

Why does this work, and why is it sometimes so spectacular? The answer probably lies with the fact that maps are much more sensitive to phases than amplitudes and, if the model bias of predicting new amplitudes is not too great, then using a nonzero value is better than zero, which is the default. Fourier-truncation errors may also be reduced, resulting in less spurious map detail.

16.1.8. Utilizing Pattersons for better starts

When slightly heavier atoms such as sulfur are present, it is possible to start recycling procedures from a set of atomic positions that are consistent with the Patterson function. For large structures, the vectors between such atoms will correspond to Patterson densities around or even below the noise level, so classical methods of locating the positions of these atoms unambiguously from the Patterson are unlikely to succeed. Nevertheless, the Patterson function can still be used to filter sets of starting atoms. This filter is currently implemented as follows in *SHELXD*. First, a sharpened Patterson function (Sheldrick *et al.*, 1993) is calculated, and the top 200 (for example) non-Harker peaks further than a given minimum distance from the origin are selected, in turn, as two-atom translation-search fragments, one such fragment being employed per solution attempt. For each of a large number of random translations, all unique Patterson vectors involving the two atoms and their symmetry equivalents are found and sorted in order of increasing Patterson density. The sum of the smallest third of these values is used as a figure of merit (PMF). Tests showed that although the globally highest PMF for a given two-atom search fragment may not correspond to correct atomic positions, nevertheless, by limiting the number of trials, some correct solutions may still be found. The two-atom vectors are chosen by biased random sampling that favours the vectors corresponding to higher Patterson values. The two atoms

Table 16.1.8.1
Overall success rates for full structure solution for hirustasin using different two-atom search vectors chosen from the Patterson peak list

Resolution (Å)	Two-atom search fragments	Solutions per 1000 attempts
1.2	Top 100 general Patterson peaks	86
1.2	Top 300 general Patterson peaks	38
1.2	One vector, error = 0.08 Å	14
1.2	One vector, error = 0.38 Å	41
1.2	One vector, error = 0.40 Å	219
1.2	One vector, error = 1.69 Å	51
1.4	Top 100 general Patterson peaks	10
1.5	Top 100 general Patterson peaks	4
1.5	One vector, error = 0.29 Å	61

may be used to generate further atoms using a full Patterson superposition minimum function or a weighted difference synthesis.

In the case of the small protein BPTI (Schneider, 1998), 15 300 attempts based on 100 different search vectors led to four final solutions with mean phase error less than 18°, although none of the globally highest PMF values for any of the search vectors corresponded to correct solutions. Table 16.1.8.1 shows the effect of using different two-atom search fragments for hirustasin, a previously unsolved 55-amino-acid protein containing five disulfide bridges first solved using *SHELXD* (Usón *et al.*, 1999). It is not clear why some search fragments perform so much better than others; surprisingly, one of the more effective search vectors deviates considerably (1.69 Å) from the nearest true S–S vector.

16.1.9. *Shake-and-Bake*: an analysis of a dual-space method in action

The *Shake-and-Bake* algorithm generated the *SnB* program written in Buffalo at the Hauptman–Woodward Institute, principally by Charles Weeks and Russ Miller (Miller *et al.*, 1994; Weeks & Miller, 1999*a*). *SHELXD* (Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002) and later *HySS* (Grosse-Kunstleve & Adams, 2003) were both inspired by *SnB* and employ the *Shake-and-Bake* philosophy with various modifications, in particular involving the use of the Patterson function to obtain starting phases. It is instructive to see how such software works in detail.

16.1.9.1. Flowchart and program comparison

A flowchart for the generic *Shake-and-Bake* algorithm, which provides the foundation for these programs, is presented in Fig. 16.1.9.1. It contains two refinement loops embedded in the trial-structure loop. The first of these loops (steps 5–9) is a dual-space phase-improvement loop entered by all trial structures, and the second (steps 11–14) is a real-space Fourier-refinement loop entered only by those trial structures that are currently judged to be the best on the basis of some figure of merit. These loops have been called the internal and external loops, respectively, in previous descriptions of the *SHELXD* program (*e.g.* Sheldrick & Gould, 1995; Sheldrick, 1997, 1998). Currently, the major algorithmic differences between the programs are the following:

(*a*) During the reciprocal-space segment of the dual-space loop (Fig. 16.1.9.1, step 5), *SnB* can perform tangent refinement or use parameter shift to reduce the minimal function [equation (16.1.4.2)] or an exponential variant of the minimal function (Hauptman *et al.*, 1999). *SHELXD* performs Karle-type tangent expansion (Karle, 1968). During tangent or parameter-shift refinement with *SnB*, all phases computed in the preceding structure-factor calculation (steps 4 or 9) are refined. During tangent expansion in *SHELXD*, the phases of (typically) the 40% highest calculated E magnitudes are held fixed, and the phases of the remaining 60% are determined by using the tangent formula. If there is a tendency for *SHELXD* to produce uranium-atom solutions, more phases should be held fixed in the tangent phase expansion.

(*b*) In real space, *SnB* uses simple peak picking, varying the number of peaks selected on the basis of structure size and composition. *SHELXD* contains provisions for all the forms of peak picking described above.

(*c*) *SnB* relies primarily on the minimal function [equation (16.1.4.2)] as a figure of merit whereas *SHELXD* uses the correlation coefficient [equation (16.1.6.1)], calculated using all data, after the final dual-space (internal) cycle and in the real-space (external) loop. In addition, *SHELXD* calculates a further correlation coefficient, $\mathrm{CC}_{\mathrm{weak}}$, calculated in the same

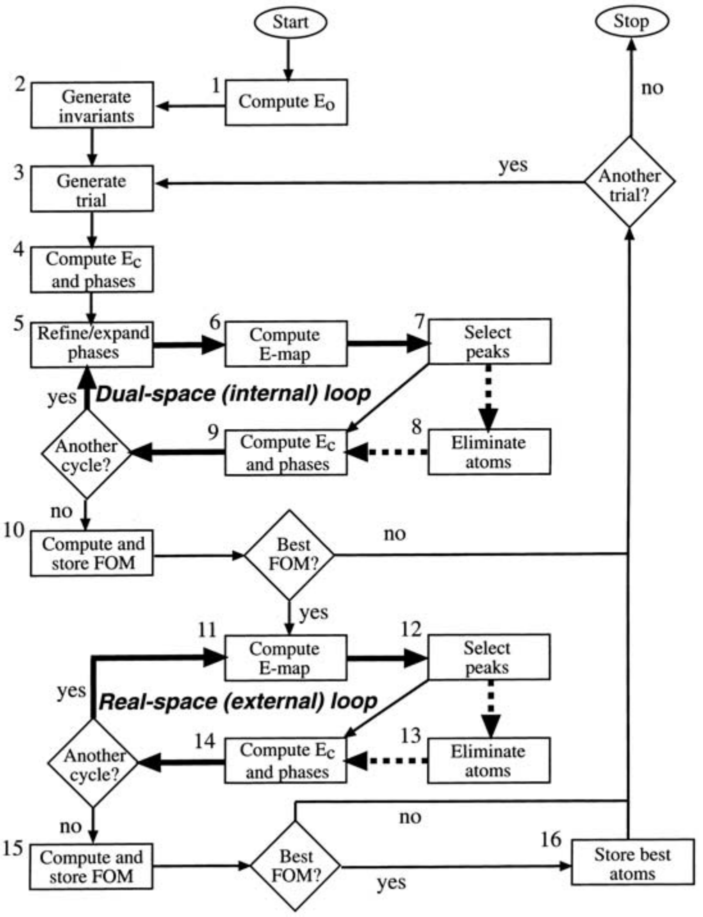


Figure 16.1.9.1
A flowchart for the *Shake-and-Bake* procedure, which is implemented in both *SnB* and *SHELXD*. The essence of the method is the dual-space approach of refining trial structures as they shuttle between real and reciprocal space. In the general case, steps 7 and 12 are any density-modification procedure, and steps 9 and 14 are inverse Fourier transforms rather than structure-factor calculations. The optional steps 8 and 13 take the form of *iterative peaklist optimization* or *random omit maps* in *SHELXD*. Any suitable starting model can be used in step 3, and *SHELXD* attempts to improve on random models (when possible) by utilizing Patterson-based information. Step 4 is bypassed if phase sets (random or otherwise) provide the starting point for the dual-space loop. *SHELXD* enters the real-space loop if the FOM (correlation coefficient) is within a specified threshold (1–5%) of the best value so far.

Table 16.1.9.1
Recommended parameter values for the *SnB* program

Values are expressed in terms of N_u, the number of unique non-H atoms (solvent atoms are typically ignored). Full-structure recommendations are for data sets measured to 1.1 Å resolution or better. Only heavy atoms or anomalous scatterers are counted for substructures.

Parameter	Full structures	Substructures
Phases	$10N_u$	$30N_u$
Triplet invariants	$100N_u$	$300N_u$
Peaks (with S, Cl) Peaks (no 'heavy')	$0.4N_u$ $0.8N_u$	N_u
Cycles	$N_u/2$ if $N_u < 100$ or if $N_u < 400$ with S, Cl *etc.*; N_u otherwise	$2N_u$ (minimum 20)

way but using only the weak reflections, *i.e.* those not used directly for phasing.

16.1.9.2. *Parameters and procedures*

All of the major parameters of the *Shake-and-Bake* procedure (*i.e.*, the numbers of refinement cycles, phases, triplet invariant relationships and peaks selected) are a function of structure size and can be expressed in terms of N_u, the number of unique non-H atoms in the asymmetric unit. These parameters have been fine-tuned in a series of tests using data for both small and large molecules (Weeks, DeTitta *et al.*, 1994; Chang *et al.*, 1997; Weeks & Miller, 1999*b*). Default (recommended) parameter values used in the *SnB* program are summarized in Table 16.1.9.1. At resolutions in the 1.1–1.4 Å range, recalcitrant data sets can sometimes be made to yield solutions if (1) the phase:invariant ratio is increased from 1:10 to values ranging between 1:20 and 1:50 or (2) the number of dual-space refinement cycles is doubled or tripled. The presence of moderately heavy atoms (*e.g.* S, C, Fe) greatly increases the probability of success at resolutions less than 1.2 Å; in general, the higher the fraction of such atoms the more the resolution requirement can be relaxed, provided that these atoms have low *B* values. Thus, disulfide bridges are much more helpful than methionine sulfur atoms because they tend to have lower *B* values. Parameter recommendations for substructures are based on an analysis of the peak-wavelength anomalous-difference data for S-adenosylhomocysteine (AdoHcy) hydrolase (Turner *et al.*, 1998). Parameter shift with a maximum of two 90° steps [indicated by the shorthand notation PS(90°, 2)] is the default phase-refinement mode. However, some structures (especially large *P*1 structures) may respond better to a single larger shift [*e.g.* PS(157.5°, 1)] (Deacon *et al.*, 1998). This seems to reduce the frequency of false minima (see Section 16.1.10.1).

In general, the parameter values used in *SHELXD* are similar to those used in *SnB*. However, the *combination* of random omit maps with tangent extension has been found to be the most effective strategy within the context of *SHELXD* for *ab initio* solution of the full stucture, and so is used as the default mode (see Section 16.1.10.2 for details). For substructure solution, especially for small substructures, it is normally faster to use starting atoms from Patterson seeding. Although both random omit and Patterson seeding can increase the success rate by an order of magnitude, combining both does not produce much further improvement. For very large substructures, and especially for very high symmetry space groups where the Patterson analysis is more time consuming, the random-omit procedure can be the more effective of the two. The largest substructure solved by *SHELXD* is probably PDB code 2pnk (to be published), solved by Qingping Xu of the Joint Center for Structural Genomics (JCSG), with 197 correct and no incorrect Se sites out of 205 (the other eight were disordered). About 1.6 million trials were needed (using multiple CPUs) to obtain one correct solution when Patterson seeding was employed, but with the random-omit method many good solutions were obtained. This example also illustrates the point that it is important not to give up too soon!

The substructure solution program *HySS* in the *PHENIX* system is more-or-less a clone of *SHELXD*. For further details see Section 16.1.12.5.

16.1.9.3. *Recognizing solutions*

On account of the intensive nature of the computations involved, *SnB* and *SHELXD* are designed to run unattended for long periods while also providing ways for the user to check the status of jobs in progress. The progress of current *SnB* jobs can be followed by monitoring a figure-of-merit histogram for the trial structures that have been processed (Fig. 16.1.9.2). A clear bimodal distribution of figure-of-merit values is a strong indication that a solution has, in fact, been found. However, not all solutions are so obvious, and it sometimes pays to inspect the best trial even when the histogram is unimodal. The course of a typical solution as a function of *SnB* cycle is contrasted with that of a nonsolution in Fig. 16.1.9.3. Minimal-function values for a solution usually decrease abruptly over the course of just a few cycles, and a tool is provided within *SnB* that allows the user to visually inspect the trace of minimal-function values for the best trial completed so far. Fig. 16.1.9.3 shows that the abrupt decrease in minimal-function values corresponds to a simultaneous abrupt increase in the number of peaks close to true atomic positions. In

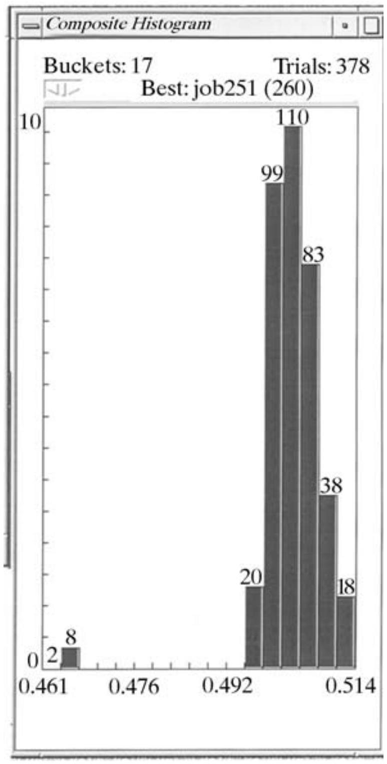


Figure 16.1.9.2
A histogram of figure-of-merit values (minimal function) for 378 scorpion toxin II trials. This bimodal histogram suggests that ten trials are solutions.

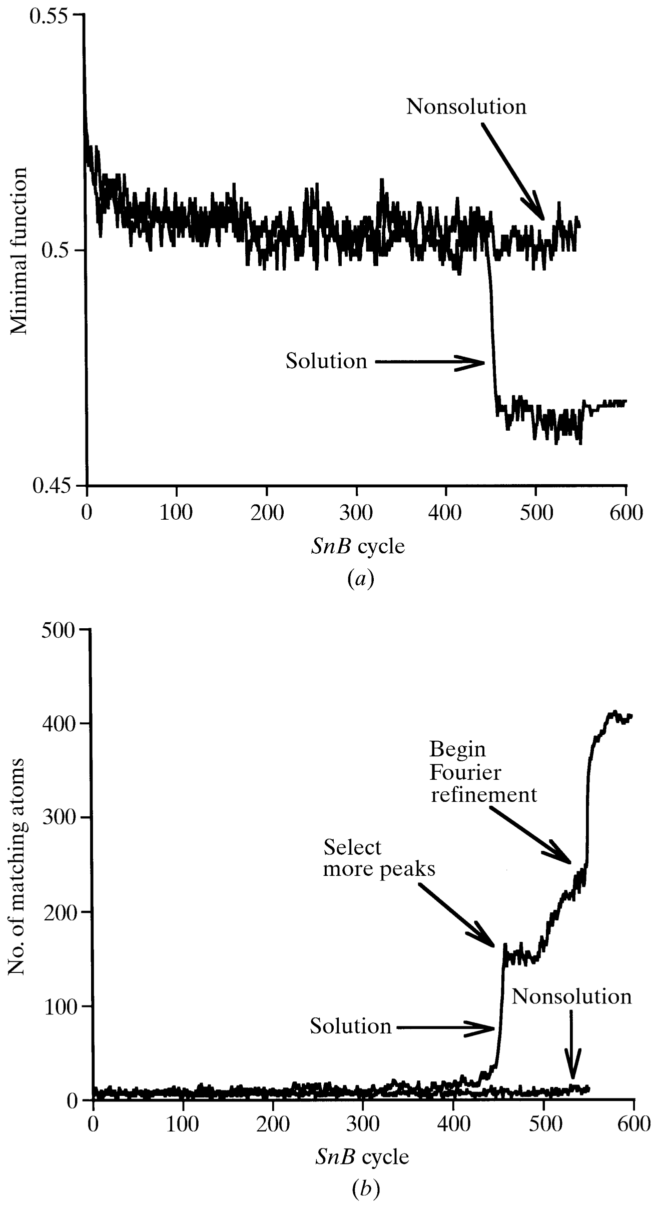


Figure 16.1.9.3

Tracing the history of a solution and a nonsolution trial for scorpion toxin II as a function *of Shake-and-Bake* cycle. (*a*) Minimal-function figure of merit, and (*b*) number of peaks closer than 0.5 Å to true atomic positions. Simple peak picking (200 or $0.4N_u$ peaks) was used for 500 (N_u) cycles, and 500 peaks (N_u) were then selected for an additional 50 ($0.1N_u$) dual-space cycles. The solution (which had the lowest minimal-function value) was then subjected to 50 cycles of Fourier refinement.

this example, a second abrupt increase in correct peaks occurs when Fourier refinement is started.

Since the correlation coefficient is a relatively absolute figure of merit (given atomic resolution, values greater than 65% almost invariably correspond to correct solutions), it is usually clear when *SHELXD* has solved a structure, although when the data do not extend to atomic resolution the CC values are less informative, and for a substructure they depend strongly on the data quality.

16.1.10. Applying dual-space programs successfully

The solution of the (known) structure of triclinic lysozyme by *SHELXD* and shortly afterwards by *SnB* (Deacon *et al.*, 1998) finally broke the 1000-atom barrier for direct methods (there happen to be 1001 protein atoms in this structure!). Both programs have also solved a large number of previously unsolved structures that had defeated conventional direct methods; some examples are listed in Table 16.1.10.1. The overall quality of solutions is generally very good, especially if appropriate action is taken during the Fourier-refinement stage. Most of the time, the *Shake-and-Bake* method works remarkably well, even for rather large structures. However, in problematic situations, the user needs to be aware of options that can increase the chance of success.

16.1.10.1. Avoiding false minima

The frequent imposition of real-space constraints appears to keep dual-space methods from producing most of the false minima that plague practitioners of conventional direct methods. Translated molecules have not been observed (so far), and traditionally problematic structures with polycyclic ring systems and long aliphatic chains are readily solved (McCourt *et al.*, 1996, 1997). False minima of the type that occur primarily in space groups lacking translational symmetry and are characterized by a single large 'uranium' peak do occur frequently in *P*1 and occasionally in other space groups. Triclinic hen egg-white lysozyme exhibits this phenomenon regardless of whether parameter-shift or tangent-formula phase refinement is employed. An example from another space group (*C*222) is provided by the Se substructure data for AdoHcy hydrolase (Turner *et al.*, 1998). In this case, many trials converge to false minima if the feature in the *SnB* program that eliminates peaks at special positions is not utilized.

The problem with false minima is most serious if they have a 'better' value of the figure of merit being used for diagnostic purposes than do the true solutions. Fortunately, this is not the case with the uranium 'solutions', which can be distinguished on the basis of the minimal function [equation (16.1.4.2)] or the correlation coefficient [equation (16.1.6.1)]. However, it would be inefficient to compute the latter in each dual-space cycle since it requires that essentially all reflections be used. To be an effective discriminator, the figure of merit must be computed using the phases calculated from the point-atom model, not from the phases directly after refinement. Phase refinement can and does produce sets of phases, such as the uranium phases, which do not correspond to physical reality. Hence, it should not be surprising that such phase sets might appear 'better' than the true phases and could lead to an erroneous choice for the best trial. Peak picking, followed by a structure-factor calculation in which the peaks are sensibly weighted, converts the phase set back to physically allowed values. If the value of the minimal function computed from the refined or *unconstrained* phases is denoted by $R_{\rm unc}$ and the value of the minimal function computed using the *constrained* phases resulting from the atomic model is denoted by $R_{\rm con}$, then a function defined by

$$R \text{ ratio} = (R_{\rm con} - R_{\rm unc})/(R_{\rm con} + R_{\rm unc}) \qquad (16.1.10.1)$$

can be used to distinguish false minima from other nonsolutions as well as the true solutions (Xu *et al.*, 2000). Once a trial falls into a false minimum, it never escapes. Therefore, the *R* ratio can be used, within *SnB*, as a criterion for early termination of unproductive trials. Based on data for several *P*1 structures, it appears that termination of trials with *R* ratio values exceeding 0.2 will eliminate most false minima without risking rejection of any potential solutions. In the case of triclinic lysozyme, false minima can be recognized, on average, by cycle 25. Since the default recommendation would be for 1000 cycles, a substantial saving in CPU time is realized by using the *R* ratio early-termination test.

Table 16.1.10.1
Some large structures solved by the *Shake-and-Bake* method

Previously known test data sets are indicated by an asterisk (*). When two numbers are given in the resolution column, the second indicates the lowest resolution at which truncated data have yielded a solution. The program codes are *SnB* (S) and *SHELXD* (D). The largest substructures solved by these two programs are mentioned in the text.

Compound	Space group	N_u (molecule)	N_u + solvent	N_u (heavy)	Resolution (Å)	Program	Reference
Hirustasin	$P4_32_12$	402	467	10S	1.2–1.55	D	[1]
Cyclodextrin derivative	$P2_1$	448	467	—	0.88	D	[2]
Alpha-1 peptide	$P1$	408	471	Cl	0.92	S	[3]
Rubredoxin*	$P2_1$	395	497	Fe, 6S	1.0–1.1	S, D	[4]
Vancomycin	$P1$	404	547	12Cl	0.97	S	[5]
BPTI*	$P2_12_12_1$	453	561	7S	1.08	D	[6]
Cyclodextrin derivative	$P2_1$	504	562	28S	1.00	D	[7]
Balhimycin*	$P2_1$	408	598	8Cl	0.96	D	[8]
Mg-complex*	$P\bar{1}$	576	608	8Mg	0.87	D	[9]
Scorpion toxin II*	$P2_12_12_1$	508	624	8S	0.96–1.2	S	[10]
Bucandin	$C2$	516	634	10S	1.05	D	[11]
Decaplanin	$P2_1$	448	635	4Cl	1.00	D	[12]
Amylose-CA26	$P1$	624	771	—	1.10	D	[13]
Viscotoxin B2	$P2_12_12_1$	722	818	12S	1.05	D	[14]
Mersacidin	$P3_2$	750	826	24S	1.04	D	[15]
Cv HiPIP H42Q*	$P2_12_12_1$	631	837	4Fe	0.93	D	[16]
Feglymycin	$P6_5$	828	1026	—	1.10	D	[17]
Acutohaemolysin	$C2_1$	1010	1242	17S	0.8	S	[18]
Tsuchimycin	$P1$	1069	1293	24Ca	1.00	D	[19]
HEW lysozyme*	$P1$	1001	1295	10S	0.85	S, D	[20], [21]
rc-WT Cv HiPIP	$P2_12_12_1$	1264	1599	8Fe	1.20	D	[16]
Cytochrome c3	$P3_1$	2024	2208	8Fe	1.20	D	[22]

References: [1] Usón *et al.* (1999); [2] Aree *et al.* (1999); [3] Privé *et al.* (1999); [4] Dauter *et al.* (1992); [5] Loll *et al.* (1998); [6] Schneider (1998); [7] Reibenspiess *et al.* (2000); [8] Schäfer *et al.* (1998); [9] Teichert (1998); [10] Smith *et al.* (1997); [11] Kuhn *et al.* (2000); [12] Lehmann *et al.* (2003); [13] Gessler *et al.* (1999); [14] Pal *et al.* (2008); [15] Schneider *et al.* (2000); [16] Parisini *et al.* (1999); [17] Bunkóczi *et al.* (2005); [18] Liu *et al.* (2003); [19] Bunkóczi (2004); [20] Deacon *et al.* (1998); [21] Walsh *et al.* (1998); [22] Frazão *et al.* (1999).

It should be noted that *SHELXD* optionally deletes the highest peak if the second peak is less than a specified fraction (*e.g.* 40%) of the height of the first, in an attempt to 'kick' the structure out of a false minimum.

Recognizing false minima is, of course, only part of the battle. It is also necessary to find a real solution, and essentially 100% of the triclinic lysozyme trials were found to be false minima when the standard parameter-shift conditions of two 90° shifts were used. In fact, significant numbers of solutions occur only when single-shift angles in the range 140–170° are used (Fig. 16.1.10.1), and there is a surprisingly high *success rate* (percentage of trial structures that go to solutions) over a narrow range of angles centred about 157.5°. It is also not surprising that there is a correlated decrease in the percentage of false minima in the range 140–150°. This suggests that a fruitful strategy for structures that exhibit a large percentage of false minima would be the following. Run 100 or so trials at each of several shift angles in the range 90–180°, find the smallest angle which gives nearly zero false minima, and then use this angle as a single shift for many trials. Balhimycin (Schäfer *et al.*, 1998) is an example of a large non-$P1$ structure that also requires a parameter shift of around 154° to obtain a solution using the minimal function.

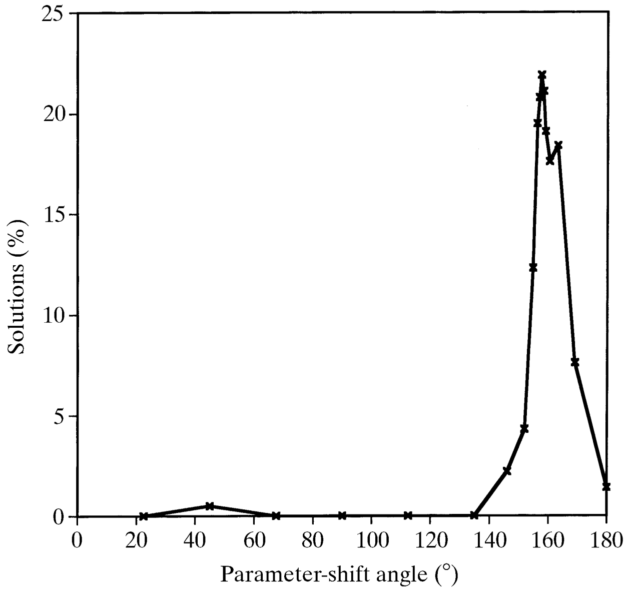


Figure 16.1.10.1
Success rates for triclinic lysozyme are strongly influenced by the size of the parameter-shift angle. Each point represents a minimum of 256 trials.

16.1.10.2. Choosing a refinement strategy

Variations in the computational details of the dual-space loop can make major differences in the efficacy of *SnB* and *SHELXD*. Fig. 16.1.10.2 shows the results of different strategies tested on a 148-atom $P1$ structure (Karle *et al.*, 1989) while developing *SHELXD*. The CPU time requirements of parameter-shift (PS) and tangent-formula expansion (TE) are similar, both being slower than no phase refinement (NR). In real space, the random-omit-map strategy (RO) was slightly faster than simple peak picking (PP) because fewer atoms were used in the structure-factor calculations. Both of these procedures were much faster than iterative peaklist optimization (PO). The original *SHELXD* algorithm (TE + PO) performs quite well in comparison with the *SnB* algorithm (PS + PP) in terms of the percentage of correct solutions, but less well when the efficiency is compared in terms of CPU time per solution. Surprising, the two strategies involving random omit maps (PS + RO and TE + RO), which had been included in the test as placebos, are much more effective than the other algorithms, especially in terms of CPU efficiency. Indeed these two runs appear to approach a 100% success rate as the number of cycles becomes large. The combination of random omit maps and Karle-type tangent

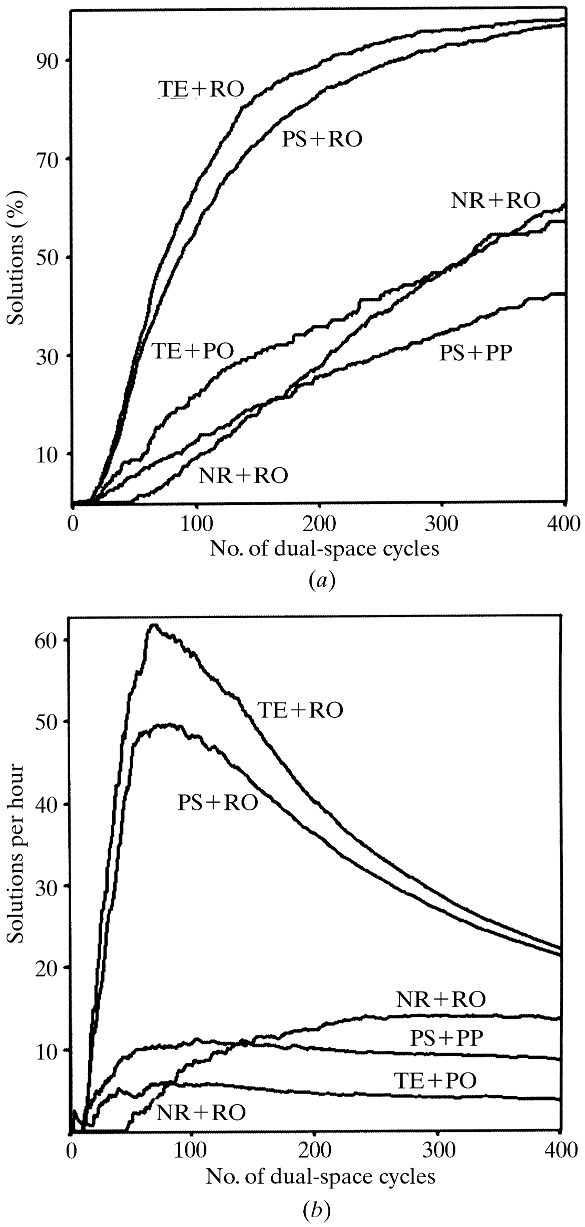


Figure 16.1.10.2

(*a*) Success rates and (*b*) cost effectiveness for several dual-space strategies as applied to a 148-atom $P1$ structure. The *phase-refinement strategies* are: (PS) parameter-shift reduction of the minimal-function value, (TE) Karle-type tangent expansion (holding the top 40% highest E_c fixed) and (NR) no phase refinement but Sim (1959) weights applied in the E map (these depend on E_c and so cannot be employed after phase refinement). The *real-space strategies* are: (PP) simple peak picking using $0.8N_u$ peaks, (PO) peaklist optimization (reducing N_u peaks to $2N_u/3$), and (RO) random omit maps (also reducing N_u peaks to $2N_u/3$). A total of about 10 000 trials of 400 internal loop cycles each were used to construct this diagram.

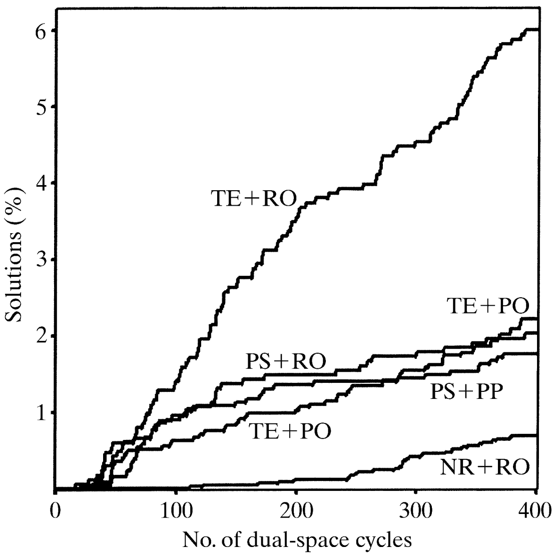


Figure 16.1.10.3

Success rates for the 317-atom $P2_12_12_1$ structure of gramicidin A.

expansion appears to be even more effective (Fig. 16.1.10.3) for gramicidin A, a $P2_12_12_1$ structure (Langs, 1988). It should be noted that conventional direct methods incorporating the tangent formula tend to perform better for this space group than in $P1$, perhaps because there is less risk of a uranium-atom pseudo-solution.

Subsequent tests using *SHELXD* on several other structures have shown that the use of random omit maps is much more effective than picking the same final number of peaks from the top of the peak list. However, it should be stressed that it is the combination TE + RO that is particularly effective. A possible special case is when a very small number of atoms is sought (*e.g.* Se atoms from MAD data). Preliminary tests indicate that peaklist optimization (PO) is competitive in such cases because the CPU time penalty associated with it is much smaller than when many atoms are involved.

With hindsight, it is possible to understand why the random omit maps provide such an efficient *search algorithm*. In macromolecular structure refinement, it is standard practice to omit parts of the model that do not fit the current electron density well, to perform some refinement or simulated annealing (Hodel *et al.*, 1992) on the rest of the model to reduce memory effects, and then to calculate a new weighted electron-density map (omit map). If the original features reappear in the new density, they were probably correct; in other cases the omit map may enable a new and better interpretation. Thus, random omit maps should not lead to the loss of an essentially correct solution, but enable efficient searching in other cases. It is also interesting to note that the results presented in Figs. 16.1.10.2 and 16.1.10.3 show that it is possible, albeit much less efficiently, to solve both structures using random omit maps without the use of any phase relationships based on probability theory (curves NR + RO).

16.1.10.3. Expansion to P1

The results shown in Table 16.1.1.1 and Fig. 16.1.10.2 indicate that success rates in space group $P1$ can be anomalously high. This suggests that it might be advantageous to expand all structures to $P1$ and then to locate the symmetry elements afterwards. However, this is more computationally expensive than performing the whole procedure in the true space group, and in practice such a strategy is only competitive in low-symmetry space groups such as $P2_1$, $C2$ or $P\bar{1}$ (Chang *et al.*, 1997). Expansion to $P1$ also offers some opportunities for starting from ‘slightly better than random’ phases. One possibility, successfully demonstrated by Sheldrick & Gould (1995), is to use a rotation search for a small fragment (*e.g.* a short piece of α-helix) to generate many sets of starting phases; after expansion to $P1$ the translational search usually required for molecular replacement is not needed. Various Patterson superposition minimum functions (Sheldrick & Gould, 1995; Pavelčík, 1994) can also provide an excellent start for phase determination for data expanded to $P1$. Drendel *et al.* (1995) were successful in solving small organic structures *ab initio* by a Fourier recycling method using perturbed Fourier amplitudes and data expanded to $P1$ without the use of probability theory. The random-omit procedure combined with expansion to $P1$ in *SHELXD* also enables structures to be solved efficiently even when the tangent formula phase extension is switched off; this has the advantage that lower E values can be used than would be suitable for the tangent formula, but at the cost of increasing the CPU time per solution. The program *ACORN2* (Dodson & Woolfson, 2009) is also particularly effective in $P1$; it applies sophisticated density modification and dual-space recycling with a special density disturbance term (POWDM) that is applied every tenth cycle. For the $P1$ form of lysozyme (see Table 16.1.10.1), good phases can be obtained by *ACORN2* starting from a fragment as small as two sulfur atoms for the 0.93 Å data. Expansion of the data to $P1$ is an essential feature of the charge-flipping approach described in Section 16.1.12.6. In general, one can say that dual-space recycling of data expanded to $P1$ requires some reasonable perturbation of the density (*e.g.* charge flipping or random peak omit) to prevent stagnation, but with this precaution provides a simple and effective approach to structure solution.

16.1.10.4. *Substructure applications*

It has been known for some time that conventional direct methods can be a valuable tool for locating the positions of heavy-atom substructures using isomorphous (Wilson, 1978) and anomalous (Mukherjee *et al.*, 1989) difference structure factors. Experience has shown that successful substructure applications are highly dependent on the accuracy of the difference magnitudes. As the technology for producing selenomethionine-substituted proteins and collecting accurate multiple-wavelength (MAD) data has improved (Hendrickson & Ogata, 1997; Smith, 1998), there has been an increased need to locate many selenium sites. For larger structures (*e.g.* more than about 30 Se atoms), automated Patterson interpretation methods can be expected to run into difficulties since the number of unique peaks to be analysed increases with the square of the number of atoms. Experimentally measured difference data are an approximation to the data for the hypothetical substructure, and it is reasonable to expect that conventional direct methods might run into difficulties sooner when applied to such data. Dual-space direct methods provide a more robust foundation for handling such data, which are often extremely noisy. Dual-space methods also have the added advantage that the expected number of Se atoms, N_u, which is usually known, can be exploited directly by picking the top N_u peaks. Successful applications require great care in data processing, especially if the $|F_A|$ values resulting from a MAD experiment are to be used.

SHELXD is frequently successfully employed with $|F_A|$ values derived from multiwavelength MAD data generated, for example, by the programs *SHELXC* (Sheldrick, 2008, 2010) or *XPREP* (Bruker AXS, Madison, WI). The decision at which resolution the data should be truncated for substructure determination is best taken on the basis of the correlation coefficients between the signed anomalous differences (Schneider & Sheldrick, 2002). On the other hand, *SnB* is normally applied separately to anomalous and dispersive differences. In many cases, both approaches lead to successful substructure solution. The real advantage of MAD data is that they provide more experimental phase information (*i.e.* better maps) and this is most important at medium to low resolution. The amount of data available for substructure problems is much larger than for full-structure problems with a comparable number of atoms to be located. Consequently, the user can afford to be stringent in eliminating data with uncertain measurements. Guidelines for rejecting uncertain data have been suggested (Smith *et al.*, 1998). Consideration should be limited to those data pairs $(|E_1|, |E_2|)$ [*i.e.*, isomorphous pairs $(|E_{\rm nat}|, |E_{\rm der}|)$ and anomalous pairs $(|E_{+\mathbf{H}}|, |E_{-\mathbf{H}}|)$] for which

$$\min\big[|E_1|/\sigma(|E_1|), |E_2|/\sigma(|E_2|)\big] \geq x_{\min} \qquad (16.1.10.2)$$

and

$$\frac{||E_1| - |E_2||}{[\sigma^2(|E_1|) + \sigma^2(|E_2|)]^{1/2}} \geq y_{\min}, \qquad (16.1.10.3)$$

where typically $x_{\min} = 3$ and $y_{\min} = 1$. The final choice of maximum resolution to be used should be based on inspection of the spherical shell averages $\langle|E_\Delta|^2\rangle_s$ *versus* $\langle s\rangle$ where $s = \sin(\theta)/\lambda$. The purpose of this precaution is to avoid spuriously large $|E_\Delta|$ values for high-resolution data pairs measured with large uncertainties due to imperfect isomorphism or general fall-off of scattering intensity with increasing scattering angle. Only those $|E_\Delta|$'s for which

$$|E_\Delta|/\sigma(|E_\Delta|) \geq z_{\min} \qquad (16.1.10.4)$$

(typically $z_{\min} = 3$) should be deemed sufficiently reliable for subsequent phasing. The probability of very large difference $|E|$'s (*e.g.* >5) is remote, and data sets that appear to have many such measurements should be examined critically for measurement errors. If a few such data remain even after the adoption of rigorous rejection criteria, it may be best to eliminate them individually. A paper by Blessing & Smith (1999) elaborates further data-selection criteria. On the other hand, it is also important that the phase:invariant ratio be maintained at 1:10 in order to ensure that the phases are overdetermined. Since the largest $|E|$'s for the substructure cell are more widely separated than they are in a true small-molecule cell, the relative number of possible triplets involving the largest reciprocal-lattice vectors may turn out to be too small. Consequently, a relatively small number of substructure phases (*e.g.* $10N_u$) may not have a sufficient number (*i.e.*, $100N_u$) of invariants. Since the number of triplets increases rapidly with the number of reflections considered, the appropriate action in such cases is to increase the number of reflections, as suggested in Table 16.1.9.1. This will typically produce the desired overdetermination.

It is rare for Se atoms to be closer to each other than 5 Å, and the application of *SnB* to AdoHcy hydolase data truncated to 4 and 5 Å has been successful. Success rates were less for lower-resolution data, but the CPU time required per trial was also reduced, primarily because much smaller Fourier grids were necessary. Consequently, there was no net increase in the CPU time needed to find a solution.

16.1.11. Substructure solution for native sulfurs and halide soaks

In the past, experimental phasing usually involved either the preparation of selenomethionine derivatives or the incorporation of heavy-metal ions by soaking crystals with a low concentration of the metal salt for several hours. The first of these methods required time in the wet lab and did not work well for all expression systems; the second had a low success rate. The improved quality of modern diffraction data collected from cryocooled crystals makes it now possible to exploit the weak anomalous signal from the native sulfur atoms or from halide ions introduced by soaking with a high concentration of a halide (iodide or bromide) for a few seconds immediately before cryocooling the crystal (Dauter *et al.*, 2000, 2001; Usón *et al.*,

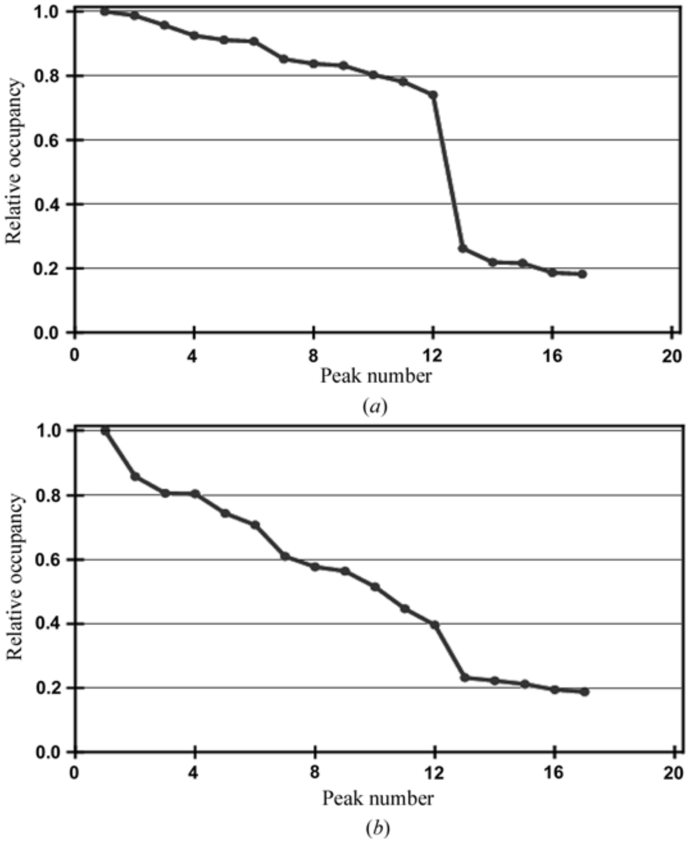


Figure 16.1.11.1

Relative occupancy against peak number for *SHELXD* substructure solutions of elastase. (*a*) Sulfur-SAD experiment showing the presence of the 12 expected sulfur atoms. (*b*) Iodide soak. Subsequent analysis showed that the peaks with relative occupancies less than 0.2 are mainly noise. These figures were made with *HKL2MAP* (Pape & Schneider, 2004).

2003). The success of these approaches is also made possible by the ability of modern, dual-space, substructure-solution programs to locate correctly a large number of sites, possibly with varying occupancies, using the SAD and SIRAS approaches.

In selenomethionine SAD and MAD phasing and in sulfur SAD phasing, the variation of the occupancies (refined in the final two cycles in the case of *SHELXD*) provides a very good indication as to whether the structure has been solved. Fig. 16.1.11.1(*a*) shows the phasing of elastase with sulfur SAD; a sharp drop in the relative occupancy after the 12th site confirms the expected presence of 12 sulfur atoms. For an iodide soak of the same protein (Fig. 16.1.11.1*b*), the relative occupancies show a gradual fall with peak number. Since the number of sites is difficult to estimate in advance for a halide soak and *SHELXD* needs to know this number approximately (within say 20%), it may be necessary to make several trials with different numbers of expected sites. From experience, the best number to use is the one that causes the occupancies to fall to about 0.2 relative to the strongest peak. Usually, subsequent refinements of the occupancies show that all the sites are partially occupied for halide soaks.

When the anomalous signal does not extend beyond about 2.0 Å, the two sulfur atoms of a disulfide bridge coalesce to a single maximum, often referred to as a supersulfur atom. At low resolution, this increases the signal-to-noise ratio for such sites in the dual-space procedure, but tends to impede phase extension to higher resolution (*e.g.* when density modification is applied to the native data with the starting phases estimated using these supersulfur atoms). An efficient way around this problem is to fit dumbbells rather than single atoms in the peak-search part of the dual-space recycling (Debreczeni *et al.*, 2003); this dramatically improves the quality of the higher-resolution starting phases.

Because the weak anomalous signal is swamped by the noise at higher resolution in such SAD experiments, it is often essential to truncate the resolution of the anomalous difference data before searching for the substructure. For MAD experiments, it is customary to truncate the data to the resolution at which the correlation coefficient between the signed anomalous differences falls below 30% (Schneider & Sheldrick, 2002). The same criterion can be used for SAD experiments if two independent data sets (*e.g.* from two different crystals) are available. As a compromise, the signed anomalous differences can be divided randomly into two sets, and then the correlation coefficient between them can be calculated. However, since these sets are not completely independent, a higher threshold (say 40%) might be advisable. An alternative criterion is to truncate the data at the point where the ratio of the mean absolute anomalous difference to its mean standard deviation falls below ~1.3, but this requires rather precise estimates of the standard deviations. In borderline cases, especially when multiple CPUs are available, it is probably safer simply to run the substructure solution for a range of different resolution cutoffs in parallel, and this is already implemented in several of the automated phasing pipelines. Sometimes good solutions are only obtained in a rather limited resolution cutoff range. A good starting value for sulfur SAD is the diffraction limit plus 0.5 Å.

16.1.12. Computer programs for dual-space phasing

Macromolecular crystallography is well served with free, high-quality, open-source software. Programs that provide direct-methods phasing for macromolecular problems will now be outlined. Although they all (except *CRUNCH2*) implement procedures that can be described more-or-less as dual-space methods, there are also appreciable differences from the three programs discussed so far. In this section, we have attempted to highlight these differences.

16.1.12.1. *ACORN*

ACORN (Yao *et al.*, 2006) and its successor *ACORN2* (Dodson & Woolfson, 2009) start with a fragment. This fragment can be very small: 1–8% in *ACORN*, and as little as 0.25% of the scattering is reported for *ACORN2*. Strictly speaking, these are not direct-methods programs, since they solve and refine crystal structures from poor starting phase sets that are usually derived from a known fragment. However, since this fragment can be very small, and since for $P1$ structures a single heavy atom at the origin suffices as a useable starting point, they are included here.

The data are normalized to give E magnitudes and partitioned into three sets: (1) large observed normalized magnitudes, (2) small magnitudes (typically < 0.2), and (3) the unobserved reflections (which are given values of unity) for a resolution range. A fragment is used to generate a set of phases, and this is followed by a sophisticated density-modification procedure:

$$\begin{aligned}
&\rho^{(n+1)} = 0 \quad \text{if} \quad \rho^{(n)} \le L\sigma,\\
&\rho^{(n+1)} = \rho^{(n)} \tanh[0.2(\rho^{(n)}/\sigma)^{\eta}] \quad \text{if} \quad \rho^{(n)} > L\sigma, \qquad (16.1.12.1)\\
&\rho^{(n+1)} = T\sigma \quad \text{if} \quad \rho^{(n+1)} > T\sigma,
\end{aligned}$$

where σ is the standard deviation of the map density and

$$T = \max(T_1 + c + 0.5c^2, 100),$$

$$T_1 = \left(\frac{M}{N}\right)^{1/2}\frac{Z_{\max}}{14}; \quad 3 \le T_1 \le 15, \qquad (16.1.12.2)$$

where M is the number of observable reflections within the resolution sphere and N is the number of atoms in the unit cell (excluding H atoms). The unconstrained value of T_1 is approximately 0.5 of the expected peak height of the heaviest atom in the E map with perfect phases; c is the cycle number.

$$L = L_1 - L_1^{c/n},$$

$$L_1 = 1.05\left[\left(B/r^2\right) - 1\right]\Phi(Z_{\max}), \qquad (16.1.12.3)$$

where B is the usual overall temperature factor; Φ is a cubic function going through the points (Φ, Z) = (0.84, 16), (0.96, 30), (1.15, 34.5) and (1.24, 48). If $Z_{\max} < 16$, the value $Z = 16$ (corresponding to sulfur) is used, and for $Z_{\max} > 48$, the value $Z = 48$ (corresponding to cadmium) is used. The value of L is thus reduced in n cycles from L_1 to zero.

$$n = \mathrm{nint}(0.5/p), \qquad (16.1.12.4)$$

where 'nint' indicates the nearest integer and

$$p = \frac{\sum_{\rm fragment} Z^2}{\sum_{\rm all\ atoms} Z^2}. \qquad (16.1.12.5)$$

Finally,

$$\eta = 17.24(r-1)^5 + 1.5. \qquad (16.1.12.6)$$

This use of η improves performance at low resolution. There is also an E limit,

$$E_{\rm lim} = [0.15 + (1/r)](0.85 + 10p), \qquad (16.1.12.7)$$

imposed with the proviso that $0.75 \le E_{\rm lim} \le 1.25$. Resolution enhancement is simple: all missing data can optionally be given a value of 1.0, which is the square root of the expectation value of E^2 from Wilson statistics. It can be seen that this procedure always needs a fragment, but this can be very small indeed. Examples include five S atoms in the catalytic domain of chitinase A1 from *Bacillus circulans* WL-12 (Matsumoto *et al.*, 1999) that account for less than 1% of the scattering density and oxidoreductase (Haynes *et al.*, 1994) with less than 0.5% of the scattering power in the initial fragment. In the latter case, the space group is *P*1 so one heavy atom can be placed at the unit-cell origin. Two atoms define the fragment: one was placed at the origin and the second in a position compatible with the Patterson map. The final mean phase error was 31.6°. *ACORN2* is available as part of the *CCP4* package at http://www.ccp4.ac.uk.

16.1.12.2. *IL MILIONE*

The *IL MILIONE* software is a product of the Bari group (Burla *et al.*, 2007). It provides a complete suite of programs for structure solution including software for processing SAD, MAD or SIRAS data or for molecular replacement. We will focus here on the *ab initio* procedures in this package. *Ab initio* phasing uses triplets and tangent procedures or Patterson methods. Direct-space refinement using density modification is employed along with the use of a resolution extension procedure. For the initial phasing, triplet invariants are evaluated by means of the P10 formula (Cascarano *et al.*, 1984). The tangent formula [equation (16.1.4.1)] is used in conjunction with these triplet-phase estimates starting with random phases and multiple starting points. An early figure of merit (eFOM) is calculated for each tangent trial and only the best trial solutions based on this are submitted to direct-space refinement. When misplaced molecular fragments are present, the structures can often be solved by the *RELAX* procedure (Burla *et al.*, 2003). In this procedure, the phases of a trial solution obtained in the correct space group are extended and refined in *P*1 by direct-space techniques. The appropriate figures of merit are used to determine the appropriate vector shift to operate on the fragment.

Patterson deconvolution may also be used in conjunction with direct methods. A superposition minimum function is used. The first peak from this procedure is always used in the phasing process. For each of the remaining peaks (the number of which depends on the size of the structure to be solved and on its data resolution), a set of phases is obtained which is ranked by a specific early figure of merit (pFOM) defined as

$$\mathrm{pFOM} = \sigma/R(\Phi), \qquad (16.1.12.8)$$

where σ is the usual standard deviation of the electron-density map and $R(\Phi)$ is the minimal function [equation (16.1.4.2)]. Irrespective of whether direct or Patterson methods are used, direct-space refinement techniques consist of cycles of electron-density modification (in which only a small fraction of the electron densities are inverted to obtain new phases), and a mask from molecular envelope calculations is applied. The correct solution is identified by a final figure of merit (fFOM), and the program automatically stops when fFOM exceeds a given threshold that depends on data resolution and structural complexity. The Patterson deconvolution methods proved to be, by far, the most efficient ones for large structures and, therefore, have been chosen as the default procedure in this case. They extended the size of the macromolecular structures that are solvable *ab initio* to more than 6000 non-H atoms in the asymmetric unit, provided that at least one calcium atom is present in the asymmetric unit and atomic resolution data are available (Burla *et al.*, 2007).

In favourable circumstances, *IL MILIONE* is also able to solve protein structures with data resolution up to 1.4–1.5 Å and to provide interpretable electron-density maps. The program has been tested using about 100 structures randomly taken from among those in the PDB with resolution better than 1.6 Å. It was able to solve all the test structures that had atomic resolution data, less than 2000 non-H atoms in the asymmetric unit (Nasym), and atoms heavier than calcium present (Zmax > 20). The solution efficiency is reduced to 84% for structures with atomic resolution data, Zmax > 20 and Nasym > 2000, and it is reduced further when these conditions are not fully met. In the presence of atoms as heavy as Ho, Au, Hg or Yb, solutions of structures composed of more than 1000 atoms have been achieved at resolutions as low as 2.0 Å. Finally, it was successful even at 1.65 Å for a case containing up to 7890 non-hydrogen atoms in the asymmetric unit (Caliandro *et al.*, 2008).

IL MILIONE can also apply direct methods to SIR and MIR data. Two different approaches may be followed for protein crystal-structure solution from isomorphous data (up to five derivatives may be used). The triplet phase invariants are estimated *via* the conditional probability distribution function,

$$P(\Phi_p|E_{p\mathbf{H}}, E_{p\mathbf{K}}, E_{p\mathbf{H+K}}, E_{d\mathbf{H}}, E_{d\mathbf{K}}, E_{d\mathbf{H+K}}) = [2\pi I_0(G\cos\Phi_p)] \qquad (16.1.12.9)$$

(Hauptman, 1982*a*; Giacovazzo *et al.*, 1988, 1996), where Φ_p is the triplet phase of the protein and

$$G = 2(\sigma^3/\sigma_2^{3/2})_p E_{p\mathbf{H}} E_{p\mathbf{K}} E_{p\mathbf{H+K}} + 2q(\sigma^3/\sigma_2^{3/2})_{\mathbf{H}} \Delta_{\mathbf{H}} \Delta_{\mathbf{K}} \Delta_{\mathbf{H+K}}$$
$$\Delta = (F_d - F_p)/\Sigma_H^{1/2}. \qquad (16.1.12.10)$$

The factor q takes into account lack of isomorphism and measurement errors, and the Δ parameters are isomorphous differences normalized with respect to the heavy-atom structure. Φ_p is expected to be close to 0 or π according to whether G is positive or negative. A starting set of phases is generated by a random process, a weighted tangent formula is applied to these phases and various trials are produced among which the correct solution may be found by a suitable figure of merit. If multiple derivatives are available, the program is able to estimate automatically, for each derivative, the scattering power of the heavy-atom structure and also to suggest which is the best derivative. The *IL MILIONE* package can be obtained from http://www.ic.cnr.it.

16.1.12.3. SHELX

The *SHELX* family of programs is widely used for small- to medium-sized structure solution and refinement. The family also contains three programs that are extensively used in macromolecular crystallography: *SHELXC*, *SHELXD* and *SHELXE*. For an overview of the *SHELX* system, see Sheldrick (2008). *SHELXC* is a housekeeping program designed to prepare the necessary files for *SHELXD* and *SHELXE*. *SHELXD* (Sheldrick, 1998; Schneider & Sheldrick, 2002; Usón & Sheldrick, 1999) is employed both for substructure solution and for *ab initio* direct methods for atomic resolution data as described elsewhere in this chapter. Fu *et al.* (2007) have shown how to adapt it to multiple CPU systems. *SHELXE* (Sheldrick, 2002, 2008, 2010) improves experimental phases from SAD, SIRAS or MAD data or starting phases from molecular replacement by iterative density modification and autotracing. *SHELX* can be obtained at http://shelx.uni-ac.gwdg.de/SHELX/.

16.1.12.4. SnB and BnP

SnB was the first program to solve small macromolecules *ab initio*, using a global cost function [equation (16.1.4.2)] that reflects how well the calculated phases fit the expected distribution of the triplets. It is fully described elsewhere in this chapter and is an effective tool in structure and substructure determination. Versions are available for multiple CPU systems (Rappleye *et al.*, 2002) and computational grids (Miller *et al.*, 2007). It is also available as part of the *BnP* package (Weeks *et al.*, 2002) that was produced in collaboration with the Biocrystallography Laboratory at the University of Pittsburgh for the experimental phasing of macromolecules. *SnB* is available from http://www.hwi.buffalo.edu/SnB/ and *BnP* from http://www.hwi.buffalo.edu/BnP/.

16.1.12.5. HySS

The substructure solution program *HySS* (Grosse-Kunstleve & Adams, 2003), which is part of the *PHENIX* package (Adams *et al.*, 2007), is closely modelled on *SHELXD* but was implemented using the cctbx libraries (Grosse-Kunstleve *et al.*, 2002). The main differences to *SHELXD* are (1) the translational search for two-atom fragments is performed by Fourier methods followed by a peak search rather than a random search, (2) the use of the tangent formula in reciprocal space is replaced by squaring the density in real space, and (3) several termination criteria are implemented so that the program can stop when the structure appears to be solved. The *PHENIX* package can be obtained at http://www.phenix-online.org/.

16.1.12.6. SUPERFLIP: charge flipping

Charge flipping is a disturbingly simple dual-space algorithm (Oszlányi & Sütő, 2004, 2005, 2008). It uses as input only the cell parameters of the structure, the reflection indices and the intensities. The intensities can be corrected for thermal motion *via* an overall temperature factor if required, and this is often beneficial. Neither chemical information nor the symmetry is explicitly used in the structure solution process. The electron density is sampled on a discrete rectangular grid of pixels with values ρ_i, $i = 1, N_{\rm pix}$. The algorithm proceeds iteratively. To begin the process, a starting set of structure factors is created by combining the experimental structure-factor amplitudes with random phases. Each iteration or cycle (numbered n) involves four steps:

(1) A trial electron density $\rho^{(n)}$ is obtained by inverse Fourier transform of the structure factors in the usual way.

(2) A modified density $g^{(n)}$ is obtained from $\rho^{(n)}$ by reversing the sign (charge flipping) of all density pixels with density below a certain positive threshold δ as follows:

$$\begin{aligned} g_i^{(n)} &= \rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} > \delta, \\ g_i^{(n)} &= -\rho_i^{(n)} \quad \text{if} \quad \rho_i^{(n)} \leq \delta. \end{aligned} \qquad (16.1.12.11)$$

(3) Modified structure factors are obtained by Fourier transform of $g^{(n)}$,

$$G_{\mathbf{H}}^{(n)} = FT[g^{(n)}]. \qquad (16.1.12.12)$$

(4) The structure factors $F_{\mathbf{H}}^{(n+1)}$ are obtained from $F_{\mathbf{H}}^{(n+1)}$ and $G_{\mathbf{H}}^{(n+1)} = |G_{\mathbf{H}}^{(n+1)}| \exp(2\pi i \varphi_{\mathbf{H}}^G)$ as follows:

   (*a*) $F_{\mathbf{H}}^{(n+1)} = |F_{\mathbf{H}}| \exp(2\pi i \varphi_{\mathbf{H}}^G)$ for $|F_{\mathbf{H}}|$ observed and strong.

   (*b*) $F_{\mathbf{H}}^{(n+1)} = |G_{\mathbf{H}}^{(n)}| \exp[2\pi i (\varphi_{\mathbf{H}}^G + 0.25)]$ for $|F_{\mathbf{H}}|$ observed and weak. In other words, for these reflections, calculated moduli are accepted unchanged and calculated phases are shifted by a constant $\Delta\varphi = \pi/2$. This means that the observed data of weak reflections are not used actively in the process, except for the knowledge that they are indeed weak. Experience shows that about 20–40% of all reflections can be considered weak. This use of the phase shifting of the weak reflections significantly improves the performance of the algorithm in cases of more complex structures (Oszlányi & Sütő, 2005); in some cases the success rate is increased by a factor of ten, in other cases a previously unsolvable structure becomes solvable by the modified algorithm.

   (*c*) $F_{\mathbf{H}}^{(n+1)} = 0$ for $|F_{\mathbf{H}}|$ unobserved.

   (*d*) $F_{\mathbf{H}}^{(n+1)} = G_{\mathbf{H}}^{(n)}$ for $\mathbf{H} = 0$. In other words, the value of F_{000} is not fixed.

The new set of structure factors enters the next cycle or iteration. The cycles are repeated until the calculation converges. Progress is monitored by a conventional R factor where a small change in R signals convergence. The parameter δ is the main variable of the iteration, and its value can be critical. It must often be determined by trial and error, but this search can be automated. The second variable parameter of the algorithm is the proportion of the reflections considered weak in each cycle.

The algorithm seeks a Fourier map that is stable against repeated flipping of all density regions below δ. Obviously, a large number of missing reflections will make the algorithm less efficient, because the missing reflections are assigned a zero amplitude, which induces large termination ripples in the Fourier map. The underlying assumption of the algorithm, that the density is close to zero in large regions of the unit cell and positive in small parts of the unit cell, is no longer fulfilled and the algorithm fails. The question of incomplete data has been addressed by Palatinus *et al.* (2007). They show that the missing data can be approximated on the basis of the Patterson map of the unknown structure optimized by the maximum-entropy method. Structures that could not be solved by the original charge-flipping algorithm can be solved in this way. For small molecules, 50% or more of the reflections can be missing, and the structure can still be reconstructed by charge flipping. The situation for macromolecules is less clear.

Symmetry is an important issue. Surprisingly, in the charge-flipping method, all structures are solved in space group *P*1, and all symmetry constraints are ignored. Attempts to impose symmetry usually damage the process fatally. The disadvantage of this is that the charge density of the whole unit cell must be determined, and not just that of the asymmetric unit. Furthermore, the symmetry elements must be located once a solution has been found. A computer program, *SUPERFLIP* (Palatinus & Chapuis, 2007), and a Java applet that demonstrates the procedure in two dimensions are freely available for download at http://escher.epfl.ch/flip/.

The charge-flipping method has been adapted to proteins (Dumas & van der Lee, 2008) and applied to a *P*1 structure with 7111 atoms [*i.e.* liver alcohol dehydogenase in complex with NADH and Cd-DMSO: 5866 protein atoms, 1241 waters and 4 Cd atoms (Meijers *et al.*, 2007)]. In common with other methods described in this chapter, charge flipping is much more effective for data to very high resolution (in this case 1.0 Å) and especially for structures containing heavier atoms. The method can also, in principle, be used for substructure determination; the solution of known substructures with as many as 120 unique Se atoms is reported in the same paper.

16.1.12.7. CRUNCH2 – Karle–Hauptman determinants

The program *CRUNCH2* is quite different to the other programs mentioned in this section. With the exception of some *E*-map recycling at the end to complete a substructure, *CRUNCH2* operates entirely in reciprocal space by maximizing higher-order Karle–Hauptman determinants (Karle & Hauptman, 1950; de Gelder *et al.*, 1993). It is incorporated into the automated *CRANK* pipeline for macromolecular structure solution (Ness *et al.*, 2004). The quality of the substructure solutions obtained appears to be at least as good as those from the dual-space programs, but it may be slower for large substructures.

16.1.13. Conclusions and the grand challenge

In practice, the main use of direct methods in macromolecular crystallography is to obtain substructures using SAD and MAD data where the limitations of the method can be relaxed. There are, of course, a few structures solved *ab initio*, but they are relatively uncommon. There is a grand challenge here: to solve *ab initio* macromolecular structures using the native data alone at resolutions more typical for macromolecules without the need for specific prior structural knowledge.

The extensive (and successful) use of atomicity constraints both in real space (peak picking) and reciprocal space (tangent formula and minimal function) make it difficult to overcome the need for atomic resolution data in the *Shake-and-Bake* methods. At lower resolution, the atomicity constraint should be replaced by another based on the recurrence of model fragments that can be predicted *a priori* from the protein sequence (*e.g.* small polyalanine α-helices, β-sheets, cofactors, bases, disulfide bridges *etc.*). The effectiveness of a very small, yet accurate, fraction of the total scattering mass in the form of a fragment or heavy atoms is apparent from the results of *ACORN2* and *IL MILIONE*.

Shortly before this chapter went to press, a paper by the Usón group (Rodríguez *et al.*, 2009) showed a possible way ahead in the case of equal-atom structures, by exploiting general features of protein secondary structure. In its current form, the method requires that the protein is at least 20% α-helical and diffracts to 2.0 Å or better, requirements that would be fulfilled by at least a quarter of the protein crystal structures deposited in the PDB. The method was successfully applied to four test structures and one previously unsolved 222 amino-acid structure that diffracted to 1.95 Å and had resisted all previous attempts at solution by molecular replacement and experimental phasing. The method exploits the power of the molecular-replacement program *PHASER* (McCoy *et al.*, 2007) to search for multiple copies of (for example) 14-residue α-helices with data truncated to 2.5 Å, retaining several thousand 'best' solutions at each stage as judged by maximum-likelihood criteria. These potential multi-helix solutions are all input into a new version of the program *SHELXE* (Sheldrick, 2010) that applies density modification and main-chain tracing iteratively. At some point, depending on the size of the structure and the quality of the data, but typically for a trial structure consisting of three or four α-helices making up some 12% of the structure, the autotracing locks in and gives a relatively complete backbone trace that can be immediately recognized both by the number of connected residues traced and a correlation coefficient between the calculated and observed *E* values. A multiple CPU computer grid is essential for performing these numerically intensive calculations in parallel, and the whole branching and pruning operation is performed under the control of the program *ARCIMBOLDO*. This approach is still at an early stage and should benefit from fine-tuning and the inevitable future increases in computer power, but it clearly has the potential to become a main-stream *ab initio* method for the solution of protein structures.

The development, in Buffalo, of the *Shake-and-Bake* algorithm and the *SnB* program has been supported by grants GM-46733 from NIH and ACI-9721373 from NSF, and computing time from the Center for Computational Research at SUNY Buffalo. HAH, CMW and RM would also like to thank the following individuals: Chun-Shi Chang, Ashley Deacon, George DeTitta, Adam Fass, Steve Gallo, Hanif Khalak, Andrew Palumbo, Jan Pevzner, Thomas Tang and Hongliang Xu, who have aided the development of *SnB*, and Steve Ealick, P. Lynne Howell, Patrick Loll, Jennifer Martin and Gil Privé, who have generously supplied data sets. The development, in Göttingen, of *SHELXD* has been supported by the BIOXHIT Consortium and the HCM Institutional Grant ERB CHBG CT 940731 from the European Commission. GMS and IU wish to thank Thammarat Aree, Gábor Bunkóczi, Zbigniew Dauter, Judit É. Debreczeni, Judith

Flippen-Anderson, Carlos Frazão, Katrin Gessler, Håkon Hope, Jörg Kärcher, Peter Kuhn, Victor Lamzin, David Langs, Christopher Lehmann, Peer Mittl, Emilio Parisini, Erich Paulus, Ehmke Pohl, Thierry Prange, Joe Reibenspiess, Martina Schäfer, Thomas Schneider, Markus Teichert, László Vértesy and Martin Walsh for discussions and/or generously providing data for structures referred to in this manuscript.

References

Adams, P. W., Afonine, P. V., Grosse-Kunstleve, R. W., Moriarty, N. W., Sauter, N. K., Zwart, P. H., Gopal, K., Ioerger, T. R., Kanbi, L., McKee, E., Pai, R. K., Hung, L.-W., Radhakannan, T., McCoy, A. J., Read, R. J., Storoni, L. C., Romo, T. D., Sacchettini, J. C. & Terwilliger, T. C. (2007). *Automated structure determination with Phenix*. In *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman. pp. 101–109. Dordrecht: Springer.

Aree, T., Usón, I., Schulz, B., Reck, G., Hoier, H., Sheldrick, G. M. & Saenger, W. (1999). *Variation of a theme: crystal structure with four octakis(2,3,6-tri-O-methyl)-gamma-cyclodextrin molecules hydrated differently by a total of 19.3 water. J. Am. Chem. Soc.* **121**, 3321–3327.

Baggio, R., Woolfson, M. M., Declercq, J.-P. & Germain, G. (1978). *On the application of phase relationships to complex structures. XVI. A random approach to structure determination. Acta Cryst.* A**34**, 883–892.

Beurskens, P. T. (1981). *A statistical interpretation of rotation and translation functions in reciprocal space. Acta Cryst.* A**37**, 426–430.

Bhuiya, A. K. & Stanley, E. (1963). *The refinement of atomic parameters by direct calculation of the minimum residual. Acta Cryst.* **16**, 981–984.

Blessing, R. H. (1997). *LOCSCL: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data. J. Appl. Cryst.* **30**, 176–177.

Blessing, R. H., Guo, D. Y. & Langs, D. A. (1996). *Statistical expectation value of the Debye–Waller factor and E(hkl) values for macromolecular crystals. Acta Cryst.* D**52**, 257–266.

Blessing, R. H. & Smith, G. D. (1999). *Difference structure-factor normalization for heavy-atom or anomalous-scattering substructure determinations. J. Appl. Cryst.* **32**, 664–670.

Bricogne, G. (1998). *Bayesian statistical viewpoint on structure determination: basic concepts and examples. Methods Enzymol.* **276**, 361–423.

Bunkóczi, G. (2004). *Structure determination of peptides with antimicrobial action*. PhD Thesis, Georg-August-Universität, Göttingen, Germany.

Bunkóczi, G., Vértesy, L. & Sheldrick, G. M. (2005). *The antiviral antibiotic feglymycin: First direct-methods solution of a 1000+ equal-atom structure. Angew. Chem. Int. Ed.* **44**, 1340–1342.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G., Siliqi, D. & Spagna, R. (2007). *IL MILIONE: a suite of computer programs for crystal structure solution of proteins. J. Appl. Cryst.* **40**, 609–613.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *SIR2004: an improved tool for crystal structure determination and refinement. J. Appl. Cryst.* **38**, 381–388.

Burla, M. C., Carrozzini, B., Caliandro, R., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Ab initio protein phasing at 1.4 Å resolution: the new phasing approach of SIR2003-N. Acta Cryst.* A**59**, 560–568.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *Ab initio phasing of proteins with heavy atoms at non-atomic resolution: pushing the size limit of solvable structures up to 7890 non-H atoms in the asymmetric unit. J. Appl. Cryst.* **41**, 548–553.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*a*). *Phasing at resolution higher than the experimental resolution. Acta Cryst.* D**61**, 556–565.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*b*). *Ab initio phasing at resolution higher than experimental resolution. Acta Cryst.* D**61**, 1080–1087.

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *The method of representations of structure seminvariants. The strengthening of triplet relationships. Acta Cryst.* A**40**, 278–283.

Chang, C.-S., Weeks, C. M., Miller, R. & Hauptman, H. A. (1997). *Incorporating tangent refinement in the Shake-and-Bake formalism. Acta Cryst.* A**53**, 436–444.

Cochran, W. (1955). *Relations between the phases of structure factors. Acta Cryst.* **8**, 473–478.

Dauter, Z. (2006). *Current state and prospects of macromolecular crystallography. Acta Cryst.* D**62**, 1–11.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Novel approach to phasing proteins: derivatization by short cryo-soaking with halides. Acta Cryst.* D**56**, 232–237.

Dauter, Z., Li, M. & Wlodawer, A. (2001). *Practical experience with the use of halides for phasing macromolecular structures: a powerful tool for structural genomics. Acta Cryst.* D**57**, 239–249.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Refinement of rubredoxin from Desulfovibrio vulgaris at 1.0 Å with and without restraints. Acta Cryst.* B**48**, 42–59.

Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *The Shake-and-Bake structure determination of triclinic lysozyme. Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.

Debaerdemaeker, T. & Woolfson, M. M. (1983). *On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement. Acta Cryst.* A**39**, 193–196.

Debaerdemaeker, T. & Woolfson, M. M. (1989). *On the application of phase relationships to complex structures. XXVIII. XMY as a random approach to the phase problem. Acta Cryst.* A**45**, 349–353.

Debreczeni, J. É., Girmann, B., Zeeck, A., Krätzner, R. & Sheldrick, G. M. (2003). *Structure of viscotoxin A3: disulfide location from weak SAD data. Acta Cryst.* D**59**, 2125–2132.

Delft, F. von, Inoue, T., Saldanha, S. A., Ottenhof, H. H., Schmitzberger, F., Birch, L. M., Dhanaraj, V., Witty, M., Smith, A. G., Blundell, T. L. & Abell, C. (2003). *Structure of E. coli ketopantoate hydroxymethyl transferase complexed with ketopantoate and Mg2+, solved by locating 160 selenomethionine sites. Structure*, **11**, 985–996.

DeTitta, G. T., Edmonds, J. W., Langs, D. A. & Hauptman, H. (1975). *Use of the negative quartet cosine invariants as a phasing figure of merit: NQEST. Acta Cryst.* A**31**, 472–479.

DeTitta, G. T., Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Structure solution by minimal-function phase refinement and Fourier filtering. I. Theoretical basis. Acta Cryst.* A**50**, 203–210.

Dodson, E. J. & Woolfson, M. M. (2009). *ACORN2: new developments of the ACORN concept. Acta Cryst.* D**65**, 881–891.

Drendel, W. B., Dave, R. D. & Jain, S. (1995). *Forced coalescence phasing: a method for ab initio determination of crystallographic phases. Proc. Natl Acad. Sci. USA*, **92**, 547–551.

Dumas, C. & van der Lee, A. (2008). *Macromolecular structure by charge flipping. Acta Cryst.* D**64**, 864–873.

Frazão, C., Sieker, L., Sheldrick, G. M., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). *Ab initio structure solution of a dimeric cytochrome c3 from Desulfovibrio gigas containing disulfide bridges. J. Biol. Inorg. Chem.* **4**, 162–165.

Fu, Z.-Q., Chrzas, J., Sheldrick, G. M., Rose, J. & Wang, B.-C. (2007). *A parallel program using SHELXD for quick heavy-atom partial structure solution on high-performance computers. J. Appl. Cryst.* **40**, 387–390.

Fujinaga, M. & Read, R. J. (1987). *Experiences with a new translation-function program. J. Appl. Cryst.* **20**, 517–521.

Gelder, R. de, de Graaff, R. A. G. & Schenk, H. (1993). *Automatic determination of crystal structures using Karle–Hauptman matrices. Acta Cryst.* A**49**, 287–293.

Germain, G., Main, P. & Woolfson, M. M. (1970). *On the application of phase relationships to complex structures. II. Getting a good start. Acta Cryst.* B**26**, 274–285.

Germain, G. & Woolfson, M. M. (1968). *On the application of phase relationships to complex structures. Acta Cryst.* B**24**, 91–96.

Gessler, K., Usón, I., Takaha, T., Krauss, N., Smith, S. M., Okada, S., Sheldrick, G. M. & Saenger, W. (1999). *V-Amylose at atomic resolution: X-ray structure of a cycloamylose with 26 glucoses. Proc. Natl Acad. Sci. USA*, **96**, 4246–4251.

Giacovazzo, C. (1976). *A probabilistic theory of the cosine invariant* $\cos(\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{\mathbf{l}} - \varphi_{\mathbf{h+k+l}})$. *Acta Cryst.* A**32**, 91–99.

Giacovazzo, C. (2008). *Direct methods.* In *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.2. Dordrecht: Kluwer Academic Publishers.

Giacovazzo, C., Cascarano, G. & Zheng, C. D. (1988). *On integrating the techniques of direct methods and isomorphous replacement.*

A new probabilistic formula for triplet invariants. *Acta Cryst.* A**44**, 45–51.

Giacovazzo, C., Siliqi, D., Platas, J. G., Hecht, H.-J., Zanotti, G. & York, B. (1996). *The ab initio crystal structure solution of proteins by direct methods. VI. Complete phasing up to derivative resolution. Acta Cryst.* D**52**, 813–825.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Substructure search procedures for macromolecular structures*. *Acta Cryst.* D**59**, 1966–1973.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework*. *J. Appl. Cryst.* **35**, 126–136.

Hauptman, H. (1974). *On the theory and estimation of the cosine invariants* $\cos(\varphi_{\rm l} + \varphi_{\rm m} + \varphi_{\rm n} + \varphi_{\rm p})$. *Acta Cryst.* A**30**, 822–829.

Hauptman, H. (1975). *A new method in the probabilistic theory of the structure invariants. Acta Cryst.* A**31**, 680–687.

Hauptman, H. (1982*a*). *On integrating the techniques of direct methods and isomorphous replacement. I. The theoretical basis. Acta Cryst.* A**38**, 289–294.

Hauptman, H. (1982*b*). *On integrating the techniques of direct methods with anomalous dispersion. I. The theoretical basis. Acta Cryst.* A**38**, 632–641.

Hauptman, H. A. (1991). *A minimal principle in the phase problem.* In *Crystallographic Computing 5: from Chemistry to Biology*, edited by D. Moras, A. D. Podjarny & J. C. Thierry, pp. 324–332. Oxford: International Union of Crystallography and Oxford University Press.

Hauptman, H. A. & Karle, J. (1953). *Solution of the phase problem. I. The centrosymmetric crystal.* Am. Crystallogr. Assoc. Monograph No. 3. Dayton, Ohio: Polycrystal Book Service.

Hauptman, H. A., Xu, H., Weeks, C. M. & Miller, R. (1999). *Exponential Shake-and-Bake: theoretical basis and applications. Acta Cryst.* A**55**, 891–900.

Haynes, M. R., Stura, E. A., Hilvert, D. & Wilson, I. A. (1994). *Crystal structures of an antibody to a peptide and its complex with peptide antigen at 2.8 Å. Science*, **263**, 646–652.

Hendrickson, W. A. & Ogata, C. M. (1997). *Phase determination from multiwavelength anomalous diffraction measurements. Methods Enzymol.* **276**, 494–523.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

Karle, I. L., Flippen-Anderson, J. L., Uma, K., Balaram, H. & Balaram, P. (1989). *α-Helix and mixed* 3_{10}*/α-helix in cocrystallized conformers of Boc-Aib-Val-Aib-Aib-Val-Val-Val-Aib-Val-Aib-Ome. Proc. Natl Acad. Sci. USA*, **86**, 765–769.

Karle, J. (1968). *Partial structural information combined with the tangent formula for noncentrosymmetric crystals. Acta Cryst.* B**24**, 182–186.

Karle, J. & Hauptman, H. (1950). *The phases and magnitudes of the structure factors*. *Acta Cryst.* **3**, 181–187.

Karle, J. & Hauptman, H. (1956). *A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22,* $3P_12$*,* $3P_22$*. Acta Cryst.* **9**, 635–651.

Kinneging, A. J. & de Graaf, R. A. G. (1984). *On the automatic extension of incomplete models by iterative Fourier calculation. J. Appl. Cryst.* **17**, 364–366.

Kuhn, P., Deacon, A. M., Comoso, S., Rajaseger, G., Kini, R. M., Usón, I. & Kolatkar, P. R. (2000). *The atomic resolution structure of bucandin, a novel toxin isolated from the Malayan krait, determined by direct methods*. *Acta Cryst.* D**56**, 1401–1407.

La Fortelle, E. de & Bricogne, G. (1997). *Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. Methods Enzymol.* **276**, 472–494.

Langs, D. A. (1988). *Three-dimensional structure at 0.86 Å of the uncomplexed form of the transmembrane ion channel peptide gramicidin A. Science*, **241**, 188–191.

Lehmann, C., Debreczeni, J. É., Bunkcóczi, G., Dauter, M., Dauter, Z., Vértesy, L. & Sheldrick, G. M. (2003). *Structures of four crystal forms of decaplanin*. *Helv. Chim. Acta*, **86**, 1478–1497.

Liu, Q., Huang, Q., Teng, M., Weeks, C. M., Jelsch, C., Zhang, R. & Niu, L. (2003). *The crystal structure of a novel, inactive, lysine 49 PLA2 from Agkistrodon acutus venom: an ultrahigh resolution, ab initio structure determination*. *J. Biol. Chem.* **278**, 41400–41408.

Loll, P. J., Miller, R., Weeks, C. M. & Axelsen, P. H. (1998). *A ligand-mediated dimerization mode for vancomycin. Chem. Biol.* **5**, 293–298.

McCourt, M. P., Ashraf, K., Miller, R., Weeks, C. M., Li, N., Pangborn, W. A. & Dorset, D. L. (1997). *X-ray crystal structures of cytotoxic oxidized cholesterols: 7-ketocholesterol and 25-hydroxycholesterol. J. Lipid Res.* **38**, 1014–1021.

McCourt, M. P., Li, N., Pangborn, W., Miller, R., Weeks, C. M. & Dorset, D. L. (1996). *Crystallography of linear molecule binary solids. X-ray structure of a cholesteryl myristate/cholesteryl pentadecanoate solid solution. J. Phys. Chem.* **100**, 9842–9847.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software*. *J. Appl. Cryst.* **40**, 658–674.

Main, P. (1976). *Recent developments in the MULTAN system – the use of molecular structure*. In *Crystallographic Computing Techniques*, edited by F. R. Ahmed, pp. 97–105. Copenhagen: Munksgaard.

Matsumoto, T., Nonata, T., Hashimoto, M., Watanabe, T. & Mitsui, Y. (1999). *Three-dimensional structure of the catalytic domain of chitinase Al from Bacillus circulars WL-12 at a very high resolution. Proc. Jpn. Acad. Ser. B*, **75**, 269–274.

Matthews, B. W. & Czerwinski, E. W. (1975). *Local scaling: a method to reduce systematic errors in isomorphous replacement and anomalous scattering measurements. Acta Cryst.* A**31**, 480–497.

Meijers, R., Hans-Werner, A., Dauter, Z., Wilson, K. S., Lamzin, V. S. & Cedergren-Zeppezauer, E. S. (2007). *Structural evidence for a ligand coordination switch in liver alcohol dehydrogenase*. *Biochemistry*, **46**, 5446–5454.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *On the application of the minimal principle to solve unknown structures. Science*, **259**, 1430–1433.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB: crystal structure determination via Shake-and-Bake. J. Appl. Cryst.* **27**, 613–621.

Miller, R., Shah, N., Green, M. L., Furey, W. & Weeks, C. M. (2007). *Shake-and-Bake on the grid*. *J. Appl. Cryst.* **40**, 938–944.

Morris, R. J., Blanc, E. & Bricogne, G. (2004). *On the interpretation and use of* $\langle|E|^2\rangle(d^*)$ *profiles*. *Acta Cryst.* D**60**, 227–240.

Morris, R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 Å rule and beyond*. *Acta Cryst.* D**59**, 615–617.

Mukherjee, A. K., Helliwell, J. R. & Main, P. (1989). *The use of MULTAN to locate the positions of anomalous scatterers. Acta Cryst.* A**45**, 715–718.

Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Crank: new methods for automated macromolecular crystal structure solution*. *Structure*, **12**, 1753–1761.

Oszlányi, G. & Sütő, A. (2004). *Ab initio structure solution by charge flipping*. *Acta Cryst.* A**60**, 134–141.

Oszlányi, G. & Sütő, A. (2005). *Ab initio structure solution by charge flipping. II. Use of weak reflections*. *Acta Cryst.* A**61**, 147–152.

Oszlányi, G. & Sütő, A. (2008). *The charge flipping algorithm*. *Acta Cryst.* A**64**, 123–134.

Pal, A., Debreczeni, J. É., Sevvana, M., Gruene, T., Kahle, B., Zeeck, A. & Sheldrick, G. M. (2008). *Structures of viscotoxins A1 and B2 from European mistletoe solved using native data alone*. *Acta Cryst.* D**64**, 985–992.

Palatinus, L. & Chapuis, G. (2007). *SUPERFLIP – a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions*. *J. Appl. Cryst.* **40**, 786–790.

Palatinus, L., Steurer, W. & Chapuis, G. (2007). *Extending the charge-flipping method towards structure solution from incomplete data sets*. *J. Appl. Cryst.* **40**, 456–462.

Pape, T. & Schneider, T. R. (2004). *HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs*. *J. Appl. Cryst.* **37**, 843–844.

Parisini, E., Capozzi, F., Lubini, P., Lamzin, V., Luchinat, C. & Sheldrick, G. M. (1999). *Ab initio solution and refinement of two high potential iron protein structures at atomic resolution. Acta Cryst.* D**55**, 1773–1784.

Pavelčík, F. (1994). *Patterson-oriented automatic structure determination. Deconvolution techniques in space group P1. Acta Cryst.* A**50**, 467–474.

Privé, G. G., Anderson, D. H., Wesson, L., Cascio, D. & Eisenberg, D. (1999). *Packed protein bilayers in the 0.9 Å resolution structure of a designed alpha helical bundle. Protein Sci.* **8**, 1400–1409.

Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *SnB version 2.2: an example of crystallographic multiprocessing. J. Appl. Cryst.* **35**, 374–376.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Refaat, L. S. & Woolfson, M. M. (1993). *Direct-space methods in phase extension and phase determination. II. Developments of low-density elimination. Acta Cryst.* D**49**, 367–371.

Reibenspiess, J. H., Maynard, D. K., Derecskei-Kovacs, A. & Vigh, G. (2000). *Crystal structures of heptakis(2,6-di-O-tert-butyldimethylsilyl)-cyclomaltoheptaose, heptakis(2-O-methyl-3,6-di-O-tert-butyldimethylsilyl)cyclomaltoheptaose and heptakis(2-O-methyl)cyclomalto heptaose*. *Carbohydrate Res.* **328**, 217–227.

Rodriguez, D. D., Grosse, C., Himmel, S., Gonzales, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Crystallographic ab initio protein structure solution below atomic resolution*. *Nat. Methods*, **6**, 651–653.

Schäfer, M., Sheldrick, G. M., Schneider, T. R. & Vértesy, L. (1998). *Structure of balhimycin and its complex with solvent molecules*. *Acta Cryst.* D**54**, 175–183.

Schenk, H. (1974). *On the use of negative quartets. Acta Cryst.* A**30**, 477–481.

Schneider, T. R. (1998). Personal communication.

Schneider, T. R., Kärcher, J., Pohl, E., Lubini, P. & Sheldrick, G. M. (2000). *Ab initio structure determination of the lantibiotic mersacidin. Acta Cryst.* D**56**, 705–713.

Schneider, T. R. & Sheldrick, G. M. (2002). *Substructure solution with SHELXD*. *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (1982). *Crystallographic algorithms for mini- and maxi-computers*. In *Computational Crystallography*, edited by D. Sayre, pp. 506–514. Oxford: Clarendon Press.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures. Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (1997). *Direct methods based on real/reciprocal space iteration*. In *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. S. Ashton, & S. Bailey, pp. 147–158. DL-CONF-97-001. Warrington: Daresbury Laboratory.

Sheldrick, G. M. (1998). *SHELX: applications to macromolecules*. In *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.

Sheldrick, G. M. (2002). *Macromolecular phasing with SHELXE*. *Z. Kristallogr.* **217**, 644–650.

Sheldrick, G. M. (2008). *A short history of SHELX*. *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2010). *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification*. *Acta Cryst.* D**66**, 479–485.

Sheldrick, G. M., Dauter, Z., Wilson, K. S., Hope, H. & Sieker, L. C. (1993). *The application of direct methods and Patterson interpretation to high-resolution native protein data. Acta Cryst.* D**49**, 18–23.

Sheldrick, G. M. & Gould, R. O. (1995). *Structure solution by iterative peaklist optimization and tangent expansion in space group P1. Acta Cryst.* B**51**, 423–431.

Shiono, M. & Woolfson, M. M. (1992). *Direct-space methods in phase extension and phase determination. I. Low-density elimination. Acta Cryst.* A**48**, 451–456.

Shmueli, U. & Wilson, A. J. C. (2008). *Statistical properties of the weighted reciprocal lattice*. *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, ch. 2.1. Dordrecht: Kluwer Academic Publishers.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetical structures. Acta Cryst.* **12**, 813–815.

Smith, G. D., Blessing, R. H., Ealick, S. E., Fontecilla-Camps, J. C., Hauptman, H. A., Housset, D., Langs, D. A. & Miller, R. (1997). *Ab initio structure determination and refinement of a scorpion protein toxin. Acta Cryst.* D**53**, 551–557.

Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. (1998). *The use of SnB to determine an anomalous scattering substructure. Acta Cryst.* D**54**, 799–804.

Smith, J. L. (1998). *Multiwavelength anomalous diffraction in macromolecular crystallography*. In *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 211–225. Dordrecht: Kluwer Academic Publishers.

Teichert, M. (1998). *Strukturaufklärung von tetranuklearen Magnesium-chelatkomplexen – Datensammlungsstrategien mit dem SMART-CCD-System*. PhD Thesis, Georg-August-Universität, Göttingen, Germany.

Turner, M. A., Yuan, C.-S., Borchardt, R. T., Hershfield, M. S., Smith, G. D. & Howell, P. L. (1998). *Structure determination of selenomethionyl S-adenosylhomocysteine hydrolase using data at a single wavelength. Nat. Struct. Biol.* **5**, 369–375.

Usón, I., Schmidt, B., von Bülow, R., Grimme, S., von Figura, K., Dauter, M., Rajashankar, K. R., Dauter, Z. & Sheldrick, G. M. (2003). *Locating the anomalous scatterer substructures in halide and sulfur phasing*. *Acta Cryst.* D**59**, 57–66.

Usón, I. & Sheldrick, G. M. (1999). *Advances in direct methods for protein crystallography*. *Curr. Opin. Struct. Biol.* **9**, 643–648.

Usón, I., Sheldrick, G. M., de La Fortelle, E., Bricogne, G., di Marco, S., Priestle, J. P., Grütter, M. G. & Mittl, P. R. E. (1999). *The 1.2 Å crystal structure of hirustasin reveals the intrinsic flexibility of a family of highly disulphide bridged inhibitors. Structure*, **7**, 55–63.

Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE. Acta Cryst.* D**63**, 1069–1074.

Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Refinement of triclinic hen egg-white lysozyme at atomic resolution. Acta Cryst.* D**54**, 522–546.

Wang, B.-C. (1985). *Solvent flattening. Methods Enzymol.* **115**, 90–112.

Weckert, E., Schwegle, W. & Hümmer, K. (1993). *Direct phasing of macromolecular structures by three-beam diffraction. Proc. R. Soc. Lond. Ser. A*, **442**, 33–46.

Weeks, C. M., Blessing, R. H., Miller, R., Mungee, R., Potter, S. A., Rappleye, J., Smith, G. D., Xu, H. & Furey, W. (2002). *Towards automated protein structure determination: BnP, the SnB-PHASES interface*. *Z. Kristallogr.* **217**, 686–693.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications. Acta Cryst.* A**50**, 210–220.

Weeks, C. M., DeTitta, G. T., Miller, R. & Hauptman, H. A. (1993). *Applications of the minimal principle to peptide structures. Acta Cryst.* D**49**, 179–181.

Weeks, C. M., Hauptman, H. A., Chang, C.-S. & Miller, R. (1994). *Structure determination by Shake-and-Bake with tangent refinement. ACA Trans. Symp.* **30**, 153–161.

Weeks, C. M. & Miller, R. (1999*a*). *The design and implementation of SnB version 2.0. J. Appl. Cryst.* **32**, 120–124.

Weeks, C. M. & Miller, R. (1999*b*). *Optimizing Shake-and-Bake for proteins. Acta Cryst.* D**55**, 492–500.

White, P. S. & Woolfson, M. M. (1975). *The application of phase relationships to complex structures. VII. Magic integers. Acta Cryst.* A**31**, 53–56.

Wilson, K. S. (1978). *The application of MULTAN to the analysis of isomorphous derivatives in protein crystallography. Acta Cryst.* B**34**, 1599–1608.

Xu, H. & Hauptman, H. A. (2004). *Statistical approach to the phase problem. Acta Cryst.* A**60**, 153–157.

Xu, H. & Hauptman, H. A. (2006). *Recent advances in direct phasing for heavy-atom substructure determination. Acta Cryst.* D**62**, 897–900.

Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. A. (2000). *Ill-conditioned Shake-and-Bake: the trap of the false minimum. Acta Cryst.* A**56**, 112–118.

Xu, H., Weeks, C. M. & Hauptman, H. A. (2005). *Optimizing statistical Shake-and-Bake for Se-atom substructure determination. Acta Cryst.* D**61**, 976–981.

Yao, J.-X. (1981). *On the application of phase relationships to complex structures. XVIII. RANTAN – random MULTAN. Acta Cryst.* A**37**, 642–644.

Yao, J. X, Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *ACORN: a review*. *Acta Cryst.* D**62**, 901–908.

International Tables for Crystallography (2012). Vol. F, Chapter 16.2, pp. 433–436.

Chapter 16.2. The maximum-entropy method

G. Bricogne

16.2.1. Introduction

The modern concept of entropy originated in the field of statistical thermodynamics, in connection with the study of large material systems in which the number of internal degrees of freedom is much greater than the number of externally controllable degrees of freedom. This concept played a central role in the process of building a quantitative picture of the multiplicity of microscopic states compatible with given macroscopic constraints, as a measure of how much remains unknown about the detailed fine structure of a system when only macroscopic quantities attached to that system are known. The collection of all such microscopic states was introduced by Gibbs under the name 'ensemble', and he deduced his entire formalism for statistical mechanics from the single premise that the equilibrium picture of a material system under given macroscopic constraints is dominated by that configuration which can be realized with the greatest combinatorial multiplicity (*i.e.* which has maximum entropy) while obeying these constraints.

The notions of ensemble and the central role of entropy remained confined to statistical mechanics for some time, then were adopted in new fields in the late 1940s. Norbert Wiener studied Brownian motion, and subsequently time series of random events, by similar methods, considering in the latter an ensemble of messages, *i.e.* 'a repertory of possible messages, and over that repertory a measure determining the probability of these messages' (Wiener, 1949). At about the same time, Shannon created information theory and formulated his fundamental theorem relating the entropy of a source of random symbols to the capacity of the channel required to transmit the ensemble of messages generated by that source with an arbitrarily small error rate (Shannon & Weaver, 1949). Finally, Jaynes (1957, 1968, 1983) realized that the scope of the principle of maximum entropy could be extended far beyond the confines of statistical mechanics or communications engineering, and could provide the basis for a general theory (and philosophy) of statistical inference and 'data processing'.

The relevance of Jaynes' ideas to probabilistic direct methods was investigated by the author (Bricogne, 1984). It was shown that there is an intimate connection between the maximum-entropy method and an enhancement of the probabilistic techniques of conventional direct methods known as the 'saddlepoint method', some aspects of which have already been dealt with in Section 1.3.4.5.2 in Chapter 1.3 of *IT* B (Bricogne, 2008).

16.2.2. The maximum-entropy principle in a general context

16.2.2.1. Sources of random symbols and the notion of source entropy

Statistical communication theory uses as its basic modelling device a discrete source of random symbols, which at discrete times $t = 1, 2, \ldots$, randomly emits a 'symbol' taken out of a finite alphabet $\mathcal{A} = \{s_i | i = 1, \ldots, n\}$. Sequences of such randomly produced symbols are called 'messages'.

An important numerical quantity associated with such a discrete source is its *entropy per symbol H*, which gives a measure of the amount of uncertainty involved in the choice of a symbol. Suppose that successive symbols are independent and that symbol i has probability q_i. Then the general requirements that H should be a continuous function of the q_i, should increase with increasing uncertainty, and should be additive for independent sources of uncertainty, suffice to define H uniquely as

$$H(q_1, \ldots, q_n) = -k \sum_{i=1}^{n} q_i \log q_i, \qquad (16.2.2.1)$$

where k is an arbitrary positive constant [Shannon & Weaver (1949), Appendix 2] whose value depends on the unit of entropy chosen. In the following we use a unit such that $k = 1$.

These definitions may be extended to the case where the alphabet $\mathcal{A}$ is a continuous space endowed with a uniform measure μ: in this case the entropy per symbol is defined as

$$H(q) = -\int_{\mathcal{A}} q(\mathbf{s}) \log q(\mathbf{s}) \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.2)$$

where q is the probability density of the distribution of symbols with respect to measure μ.

16.2.2.2. The meaning of entropy: Shannon's theorems

Two important theorems [Shannon & Weaver (1949), Appendix 3] provide a more intuitive grasp of the meaning and importance of entropy:

(1) H is approximately the logarithm of the reciprocal probability of a typical long message, divided by the number of symbols in the message; and

(2) H gives the rate of growth, with increasing message length, of the logarithm of the number of reasonably probable messages, regardless of the precise meaning given to the criterion of being 'reasonably probable'.

The entropy H of a source is thus a direct measure of the strength of the restrictions placed on the permissible messages by the distribution of probabilities over the symbols, lower entropy being synonymous with greater restrictions. In the two cases above, the maximum values of the entropy $H_{\max} = \log n$ and $H_{\max} = \log \mu(\mathcal{A})$ are reached when all the symbols are equally probable, *i.e.* when q is a uniform probability distribution over the symbols. When this distribution is not uniform, the usage of the different symbols is biased away from this maximum freedom, and the entropy of the source is lower; by Shannon's theorem (2), the number of 'reasonably probable' messages of a given length emanating from the source decreases accordingly.

The quantity that measures most directly the strength of the restrictions introduced by the non-uniformity of q is the difference $H(q) - H_{\max}$, since the proportion of N-atom random structures which remain 'reasonably probable' in the ensemble of the corresponding source is $\exp\{N[H(q) - H_{\max}]\}$. This difference may be written (using continuous rather than discrete distributions)

$$H(q) - H_{\max} = -\int_{\mathcal{A}} q(\mathbf{s}) \log[q(\mathbf{s})/m(\mathbf{s})] \, \mathrm{d}\mu(\mathbf{s}), \qquad (16.2.2.3)$$

Copyright © 2012 International Union of Crystallography

where $m(\mathbf{s})$ is the uniform distribution which is such that $H(m) = H_{\max} = \log \mu(\mathcal{A})$.

16.2.2.3. Jaynes' maximum-entropy principle

From the fundamental theorems just stated, which may be recognized as Gibbs' argument in a different guise, Jaynes' own maximum-entropy argument proceeds with striking lucidity and constructive simplicity, along the following lines:

(1) experimental observation of, or 'data acquisition' on, a given system enables us to progress from an initial state of uncertainty to a state of lesser uncertainty about that system;
(2) uncertainty reflects the existence of numerous possibilities of accounting for the available data, viewed as constraints, in terms of a physical model of the internal degrees of freedom of the system;
(3) new data, viewed as new constraints, reduce the range of these possibilities;
(4) conversely, any step in our treatment of the data that would further reduce that range of possibilities amounts to applying extra constraints (even if we do not know what they are) which are not warranted by the available data;
(5) hence Jaynes's rule: '*The probability assignment over the range of possibilities* [*i.e.* our picture of residual uncertainty] *shall be the one with maximum entropy consistent with the available data, so as to remain maximally non-committal with respect to the missing data*'.

The only requirement for this analysis to be applicable is that the 'ranges of possibilities' to which it refers should be representable (or well approximated) by ensembles of abstract messages emanating from a random source. The entropy to be maximized is then the entropy per symbol of that source.

The final form of the maximum-entropy criterion is thus that $q(\mathbf{s})$ should be chosen so as to maximize, under the constraints expressing the knowledge of newly acquired data, its entropy

$$\mathcal{S}_m(q) = -\int_V q(\mathbf{s}) \log[q(\mathbf{s})/m(\mathbf{s})]\, \mathrm{d}\mu(\mathbf{s}) \qquad (16.2.2.4)$$

relative to the 'prior prejudice' $m(\mathbf{s})$ which maximizes H in the absence of these data.

16.2.2.4. Jaynes' maximum-entropy formalism

Jaynes (1957) solved the problem of explicitly determining such maximum-entropy distributions in the case of general linear constraints, using an analytical apparatus first exploited by Gibbs in statistical mechanics.

The maximum-entropy distribution $q^{\rm ME}(\mathbf{s})$, under the prior prejudice $m(\mathbf{s})$, satisfying the linear constraint equations

$$\mathcal{C}_j(q) \equiv \int_{\mathcal{A}} q(\mathbf{s}) C_j(\mathbf{s})\, \mathrm{d}\mu(\mathbf{s}) = c_j \quad (j = 1, 2, \ldots, M), \qquad (16.2.2.5)$$

where the $\mathcal{C}_j(q)$ are linear *constraint functionals* defined by given *constraint functions* $C_j(\mathbf{s})$, and the c_j are given *constraint values*, is obtained by maximizing with respect to q the relative entropy defined by equation (16.2.2.4). An extra constraint is the normalization condition

$$\mathcal{C}_0(q) \equiv \int_{\mathcal{A}} q(\mathbf{s})\ 1\ \mathrm{d}\mu(\mathbf{s}) = 1, \qquad (16.2.2.6)$$

to which it is convenient to give the label $j = 0$, so that it can be handled together with the others by putting $C_0(\mathbf{s}) = 1$, $c_0 = 1$.

By a standard variational argument, this constrained maximization is equivalent to the unconstrained maximization of the functional

$$\mathcal{S}_m(q) + \sum_{j=0}^{M} \lambda_j \mathcal{C}_j(q), \qquad (16.2.2.7)$$

where the λ_j are Lagrange multipliers whose values may be determined from the constraints. This new variational problem is readily solved: if $q(\mathbf{s})$ is varied to $q(\mathbf{s}) + \delta q(\mathbf{s})$, the resulting variations in the functionals $\mathcal{S}_m$ and $\mathcal{C}_j$ will be

$$\begin{aligned} \delta\mathcal{S}_m &= \int_{\mathcal{A}} \{-1 - \log[q(\mathbf{s})/m(\mathbf{s})]\}\, \delta q(\mathbf{s})\, \mathrm{d}\mu(\mathbf{s}) \quad \text{and} \\ \delta\mathcal{C}_j &= \int_{\mathcal{A}} \{C_j(\mathbf{s})\}\, \delta q(\mathbf{s})\, \mathrm{d}\mu(\mathbf{s}), \end{aligned} \qquad (16.2.2.8)$$

respectively. If the variation of the functional (16.2.2.7) is to vanish for arbitrary variations $\delta q(\mathbf{s})$, the integrand in the expression for that variation from (16.2.2.8) must vanish identically. Therefore the maximum-entropy density distribution $q^{\rm ME}(\mathbf{s})$ satisfies the relation

$$-1 - \log[q(\mathbf{s})/m(\mathbf{s})] + \sum_{j=0}^{M} \lambda_j C_j(\mathbf{s}) = 0 \qquad (16.2.2.9)$$

and hence

$$q^{\rm ME}(\mathbf{s}) = m(\mathbf{s}) \exp(\lambda_0 - 1) \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right]. \qquad (16.2.2.10)$$

It is convenient now to separate the multiplier λ_0 associated with the normalization constraint by putting

$$\lambda_0 - 1 = -\log Z, \qquad (16.2.2.11)$$

where Z is a function of the other multipliers $\lambda_1, \ldots, \lambda_M$. The final expression for $q^{\rm ME}(\mathbf{s})$ is thus

$$q^{\rm ME}(\mathbf{s}) = \frac{m(\mathbf{s})}{Z(\lambda_1, \ldots, \lambda_M)} \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right]. \qquad (\text{ME1})$$

The values of Z and of $\lambda_1, \ldots, \lambda_M$ may now be determined by solving the initial constraint equations. The normalization condition demands that

$$Z(\lambda_1, \ldots, \lambda_M) = \int_{\mathcal{A}} m(\mathbf{s}) \exp\left[\sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right] \mathrm{d}\mu(\mathbf{s}). \qquad (\text{ME2})$$

The generic constraint equations (16.2.2.5) determine $\lambda_1, \ldots, \lambda_M$ by the conditions that

$$\int_{\mathcal{A}} [m(\mathbf{s})/Z] \exp\left[\sum_{k=1}^{M} \lambda_k C_k(\mathbf{s})\right] C_j(\mathbf{s})\, \mathrm{d}\mu(\mathbf{s}) = c_j \qquad (16.2.2.12)$$

for $j = 1, 2, \ldots, M$. But, by Leibniz's rule of differentiation under the integral sign, these equations may be written in the compact form

$$\frac{\partial(\log Z)}{\partial\lambda_j} = c_j \quad (j = 1, 2, \ldots, M). \qquad (\text{ME3})$$

Equations (ME1), (ME2) and (ME3) constitute the *maximum-entropy equations*.

The maximal value attained by the entropy is readily found:

$$\begin{aligned} \mathcal{S}_m(q^{\rm ME}) &= -\int_{\mathcal{A}} q^{\rm ME}(\mathbf{s}) \log\big[q^{\rm ME}(\mathbf{s})/m(\mathbf{s})\big]\, \mathrm{d}\mu(\mathbf{s}) \\ &= -\int_{\mathcal{A}} q^{\rm ME}(\mathbf{s}) \left[-\log Z + \sum_{j=1}^{M} \lambda_j C_j(\mathbf{s})\right] \mathrm{d}\mu(\mathbf{s}), \end{aligned}$$

i.e. using the constraint equations

$$\mathcal{S}_m(q^{\rm ME}) = \log Z - \sum_{j=1}^{M} \lambda_j c_j. \qquad (16.2.2.13)$$

The latter expression may be rewritten, by means of equations (ME3), as

$$\mathcal{S}_m(q^{\rm ME}) = \log Z - \sum_{j=1}^{M} \lambda_j \frac{\partial(\log Z)}{\partial \lambda_j}, \qquad (16.2.2.14)$$

which shows that, in their dependence on the λ's, the entropy and log Z are related by Legendre duality.

Jaynes' theory relates this maximal value of the entropy to the prior probability $\mathcal{P}(\mathbf{c})$ of the vector **c** of simultaneous constraint values, *i.e.* to the size of the sub-ensemble of messages of length N that fulfil the constraints embodied in (16.2.2.5), relative to the size of the ensemble of messages of the same length when the source operates with the symbol probability distribution given by the prior prejudice m. Indeed, it is a straightforward consequence of Shannon's second theorem (Section 16.2.2) as expressed in equation (16.2.2.3) that

$$\mathcal{P}^{\rm ME}(\mathbf{c}) \propto \exp(\mathcal{S}), \qquad (16.2.2.15)$$

where

$$\mathcal{S} = \log Z^N - \lambda \cdot \mathbf{c} = N\mathcal{S}_m(q^{\rm ME}) \qquad (16.2.2.16)$$

is the total entropy for N symbols.

16.2.3. Adaptation to crystallography

16.2.3.1. *The random-atom model*

The standard setting of probabilistic direct methods (Hauptman & Karle, 1953; Bertaut, 1955*a*,*b*; Klug, 1958) uses implicitly as its starting point a source of random atomic positions. This can be described in the terms introduced in Section 16.2.2.1 by using a continuous alphabet $\mathcal{A}$ whose symbols **s** are fractional coordinates **x** in the asymmetric unit of the crystal, the uniform measure μ being the ordinary Lebesgue measure $\mathrm{d}^3\mathbf{x}$. A message of length N generated by that source is then a random N-equal-atom structure.

16.2.3.2. *Conventional direct methods and their limitations*

The traditional theory of direct methods assumes a *uniform* distribution $q(\mathbf{x})$ of random atoms and proceeds to derive joint distributions of structure factors belonging to an N-atom random structure, using the asymptotic expansions of Gram–Charlier and Edgeworth. These methods have been described in Section 1.3.4.5.2.2 of *IT* B (Bricogne, 2008) as examples of applications of Fourier transforms. The reader is invited to consult this section for terminology and notation. These joint distributions of complex structure factors are subsequently used to derive conditional distributions of phases when the amplitudes are assigned their observed values, or of a subset of complex structure factors when the others are assigned certain values. In both cases, the *largest* structure-factor amplitudes are used as the conditioning information.

It was pointed out by the author (Bricogne, 1984) that this procedure can be problematic, as the Gram–Charlier and Edgeworth expansions have good convergence properties only in the vicinity of the expectation values of each structure factor: as the atoms are assumed to be uniformly distributed, these series afford an adequate approximation for the joint distribution $\mathcal{P}(\mathbf{F})$ only near the origin of structure-factor space, *i.e.* for *small* values of all the structure amplitudes. It is therefore incorrect to use these local approximations to $\mathcal{P}(\mathbf{F})$ near $\mathbf{F} = \mathbf{0}$ as if they were the global functional form for that function 'in the large' when forming conditional probability distributions involving large amplitudes.

16.2.3.3. *The notion of recentring and the maximum-entropy criterion*

These limitations can be overcome by recognizing that, if the locus $\mathcal{T}$ (a high-dimensional torus) defined by the large structure-factor amplitudes to be used in the conditioning data is too extended in structure-factor space for a single asymptotic expansion of $\mathcal{P}(\mathbf{F})$ to be accurate everywhere on it, then $\mathcal{T}$ should be broken up into sub-regions, and different local approximations to $\mathcal{P}(\mathbf{F})$ should be constructed in each of them. Each of these sub-regions will consist of a 'patch' of $\mathcal{T}$ surrounding a point $\mathbf{F}^* \neq \mathbf{0}$ located on $\mathcal{T}$. Such a point $\mathbf{F}^*$ is obtained by assigning 'trial' phase values to the known moduli, but these trial values do not necessarily have to be viewed as 'serious' assumptions concerning the true values of the phases: rather, they should be thought of as pointing to a patch of $\mathcal{T}$ and to a specialized asymptotic expansion of $\mathcal{P}(\mathbf{F})$ designed to be the most accurate approximation possible to $\mathcal{P}(\mathbf{F})$ on that patch. With a sufficiently rich collection of such constructs, $\mathcal{P}(\mathbf{F})$ can be accurately calculated anywhere on $\mathcal{T}$.

These considerations lead to the notion of *recentring*. Recentring the usual Gram–Charlier or Edgeworth asymptotic expansion for $\mathcal{P}(\mathbf{F})$ away from $\mathbf{F} = \mathbf{0}$, by making trial phase assignments that define a point $\mathbf{F}^*$ on $\mathcal{T}$, is equivalent to using a non-uniform prior distribution of atoms $q(\mathbf{x})$, reproducing the individual components of $\mathbf{F}^*$ among its Fourier coefficients. The latter constraint leaves $q(\mathbf{x})$ highly indeterminate, but Jaynes' argument given in Section 16.2.2.3 shows that there is a uniquely defined 'best' choice for it: it is that distribution $q^{\rm ME}(\mathbf{x})$ having *maximum entropy* relative to a uniform prior prejudice $m(\mathbf{x})$, and having the corresponding values $\mathbf{U}^*$ of the unitary structure factors for its Fourier coefficients. This distribution has the unique property that it rules out as few random structures as possible on the basis of the limited information available in $\mathbf{F}^*$.

In terms of the statistical mechanical language used in Section 16.2.1, the trial structure-factor values $\mathbf{F}^*$ used as constraints would be the macroscopic quantities that can be controlled externally; while the $3N$ atomic coordinates would be the internal degrees of freedom of the system, whose entropy should be a maximum under these macroscopic constraints.

16.2.3.4. *The crystallographic maximum-entropy formalism*

It is possible to solve explicitly the maximum-entropy equations (ME1) to (ME3) derived in Section 16.2.2.4 for the crystallographic case that has motivated this study, *i.e.* for the purpose of constructing $q^{\rm ME}(\mathbf{x})$ from the knowledge of a set of trial structure-factor values $\mathbf{F}^*$. These derivations are given in §3.4 and §3.5 of Bricogne (1984). Extensive relations with the algebraic formalism of traditional direct methods are exhibited in §4, and connections with the theory of determinantal inequalities and with the maximum-determinant rule of Tsoucaris (1970) are studied in §6, of the same paper. The reader interested in these topics is invited to consult this paper, as space limitations preclude their discussion in the present chapter.

16.2.3.5. *Connection with the saddlepoint method*

The saddlepoint method constitutes an alternative approach to the problem of evaluating the joint probability $\mathcal{P}(\mathbf{F}^*)$ of structure factors when some of the moduli in $\mathbf{F}^*$ are large. It is shown in §5 of Bricogne (1984), and in more detail in Section 1.3.4.5.2.2 of Chapter 1.3 of *IT* B (Bricogne, 2008), that there is complete equivalence between the maximum-entropy approach to the phase problem and the classical probabilistic approach by the method of joint distributions, provided the latter is enhanced by the adoption of the saddlepoint approximation.

References

Bertaut, E. F. (1955*a*). *La méthode statistique en cristallographie. I. Acta Cryst.* **8**, 537–543.

Bertaut, E. F. (1955*b*). *La méthode statistique en cristallographie. II. Quelques applications. Acta Cryst.* **8**, 544–548.

Bricogne, G. (1984). *Maximum entropy and the foundations of direct methods. Acta Cryst.* A**40**, 410–445.

Bricogne, G. (2008). *Fourier transforms in crystallography: theory, algorithms and applications*. In *International Tables for Crystallography*, Vol. B. *Reciprocal Space*, edited by U. Shmueli, 3rd ed., ch. 1.3. Heidelberg: Springer.

Hauptman, H. & Karle, J. (1953). *The Solution of the Phase Problem: I. The Centrosymmetric Crystal.* ACA Monograph No. 3. Pittsburgh: Polycrystal Book Service.

Jaynes, E. T. (1957). *Information theory and statistical mechanics. Phys. Rev.* **106**, 620–630.

Jaynes, E. T. (1968). *Prior probabilities. IEEE Trans. SSC*, **4**, 227–241.

Jaynes, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics.* Dordrecht: Reidel.

Klug, A. (1958). *Joint probability distribution of structure factors and the phase problem. Acta Cryst.* **11**, 515–543.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Tsoucaris, G. (1970). *A new method of phase determination. The 'maximum determinant rule'. Acta Cryst.* A**26**, 492–499.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series.* Cambridge, MA: MIT Press.

International Tables for Crystallography (2012). Vol. F, Chapter 16.3, pp. 437–442.

Chapter 16.3. *Ab initio* phasing of low-resolution Fourier syntheses

V. Y. Lunin, A. G. Urzhumtsev and A. Podjarny

16.3.1. Introduction

Low-resolution macromolecular structural information may be obtained for non-crystalline macromolecular objects by electron microscopy (see Chapter 19.6 and references therein) or small-angle X-ray scattering (see Chapter 19.3 and references therein), or by using X-ray free-electron lasers (Chapman *et al.*, 2011; Seibert *et al.*, 2011) and iterative density-reconstruction techniques (Sayre, 2008). While giving similar structural information, low-resolution crystallographic images are a more natural starting point on the way towards atomic models. The need to work at low resolution may be due to the limited diffraction power of the crystals or the failure of standard phasing methods.

A number of techniques have been suggested to estimate the phases of structure factors using various complementary sources of information, specific to a given crystal:

(i) Phasing by isomorphous replacement or anomalous scattering techniques using heavy atoms or their clusters may lead to good results, ranging from the pioneering structures reported at nearly 6 Å resolution (Green *et al.*, 1954; Perutz *et al.*, 1960; Black *et al.*, 1962) to viruses (for example, Harrison & Jack, 1975) and ribosomes (Ban *et al.*, 1998). However, working at low resolution requires special effort, high data quality and several sets of structure-factor magnitudes.

(ii) Molecular-replacement-type low-resolution searches can be carried out using conventional models, more or less coarsely (Valegård *et al.*, 1991; Jamrog *et al.*, 2003), with simplified models or with molecular envelopes (Jack *et al.*, 1975; Rayment *et al.*, 1982; Urzhumtsev & Podjarny, 1995; Hao, 2006). The simplified models, roughly reproducing molecular shapes obtained previously, may be a spherical shell for viruses (for example Johnson *et al.*, 1976; Chapman *et al.*, 1992), a sphere or a cylindrical shell for proteins and their complexes (Podjarny *et al.*, 1987; Harris, 1995; Andersson & Hovmöller, 1996; Lunin *et al.*, 2001), or several cylinders for α-helical proteins (Kalinin, 1980; Strop *et al.*, 2007).

(iii) Phasing techniques can be based on the different average diffraction power of the proteins, nucleic acids and bulk solvent, naturally or artificially introduced (Bragg & Perutz, 1952; Roth *et al.*, 1984; Carter *et al.*, 1990; Fourme *et al.*, 1995; Shepard *et al.*, 2000). These techniques require several crystals and their corresponding diffraction data sets.

(iv) In some cases, the classical direct phasing methods can be formally tried to obtain a molecular envelope at low resolution (Podjarny *et al.*, 1981; Carter *et al.*, 1990). Similarly, the phase triplets obtained experimentally by a three-beam diffraction experiment lead in some cases to relatively low-resolution images (Hölzer *et al.*, 2000).

In this chapter we discuss only *ab initio* phasing methods. By this term we mean a mathematical or computational procedure to estimate the values of the structure-factor phases using a single set of structure-factor magnitudes and only a general type of information, in contrast with most of the methods mentioned above. Low-resolution *ab initio* phasing estimates the phase values of the structure factors of the lowest resolution for a given crystal, from several tens to several hundreds of reflections in total. Depending on the size of the unit cell, a low-resolution Fourier synthesis calculated with these structure factors can show molecular positions in the unit cell, molecular envelopes (at a resolution $d_{\min} \simeq 15$ Å or lower) or secondary-structure elements (when $d_{\min} \simeq 5$–6 Å).

16.3.2. General features of low-resolution images

Low-resolution *ab initio* phasing is usually a laborious procedure. Some features of low-resolution Fourier syntheses increase the difficulties, as follows:

(i) There is a common belief that low-resolution Fourier syntheses represent molecular envelopes when the cutoff level is relatively low and the centres of the molecules when this level is high. In practice, density peaks are often shifted from molecular centres toward regions of close intermolecular contact. Bulk-solvent correction decreases the contrast between two neighbouring peaks and can result in their merging. Also, due to the relatively small number of reflections used in the calculation, a small change in phase can significantly modify the image and the positions of the peaks.

(ii) Low-resolution envelopes cannot represent sharp molecular features accurately enough and thus cannot cover all macromolecular atoms, even when the synthesis is calculated with the exact structure-factor values.

(iii) An increase in the resolution of Fourier maps does not always help to interpret them. For example, an increase in resolution from 16–25 Å to 10–12 Å often makes maps even less suitable for visual inspection, since they stop showing molecular envelope features but do not reveal secondary-structure elements.

(iv) Usually, visual inspection of low-resolution maps does not allow one to choose the correct enantiomer. In addition, the overall features of the flipped map $-\rho(\mathbf{r})$ at very low resolution are often very similar to those of the direct map $\rho(\mathbf{r})$. This might complicate the choice of the correct sign of the map.

(v) Low-resolution Fourier maps are very sensitive to missing reflections, even if their number is small.

(vi) Maps calculated with too few reflections may show a superposition of images corresponding to different choices of the origin.

(vii) The determination of twinning and the correct choice of space group are especially complicated at low resolution.

16.3.3. Low-resolution phasing

If additional experimental information cannot be used to solve the phase problem, then the search for the correct structure has to be based on some general features of the true phase set. Such features can be formulated as a selection criterion (‘score

Copyright © 2012 International Union of Crystallography

function', 'figure of merit' *etc.*) that may have a qualitative or a quantitative form (Gilmore, 2000; Lunin, Lunina, Petrova *et al.*, 2000). In the following, we use the term 'selection criterion' to reflect a quality of the phase set as a whole, and we reserve the term 'figure of merit' to estimate the accuracy of an individual phase.

Unfortunately, no selection criterion suggested so far identifies the true phase set unambiguously (Lunin, Lunina *et al.*, 2002). Therefore, the phase problem cannot be solved by a simple search for the best value of a selection criterion. Nevertheless, it has been shown that many criteria improve the ensemble of phase sets by retaining good phase sets, *i.e.* those close to the true values. In other words, when the starting ensemble of phase sets (*e.g.* randomly generated) is reduced to those phase sets with a reasonable value of a selection criterion, the percentage of good phase sets increases. Therefore, a multifiltering cyclic phasing procedure was suggested (Lunin *et al.*, 1990):

(i) A set of reflections $S_{\rm work}$ is chosen for phasing.
(ii) A large number of random phase sets are generated either directly or using models.
(iii) The generated ensemble of phase sets is filtered by applying different selection rules.
(iv) The phase sets that fit all selection rules (the selected phase sets) form the output of the cycle.

The output is used to produce estimates of individual phase values (averaged or median) and to modify the mode of generation of random phases in the next cycle. The cycle is repeated several times, varying the phase-generating mode and the selection rules, extending the set of reflections $S_{\rm work}$ to be phased *etc.*

A key feature of this procedure is that at every cycle we do not search for the phase set with the best possible fit to the selection rules, but we filter out obviously poor phase sets.

16.3.4. Phase generation and selection

16.3.4.1. *Generation of the phase sets*

This step defines how the configuration space of all phase sets is explored. First, all phases may be considered as independent variables. At the beginning of phasing when no information is available, all values are considered equally probable. When some phase information for a reflection is available, for example in the form of a probability distribution, it may be used to generate new random phases, taking into account the restrictions of the space-group symmetry. Second, the phases may be calculated through models. For example, a phase set may be calculated from a model composed of a small number of large Gaussian spheres ('globs'), the coordinates of which are now considered as primary random variables (Lunin *et al.*, 1995; Guo *et al.*, 2000). The sphere coordinates may be generated uniformly in the whole unit cell at the very beginning of phasing and inside a molecular mask later, when some phase information becomes available. The use of models parameterizes the phase space and thus reduces the search. At the same time, if the models are inappropriate, they cannot generate a good approximation to the correct solution.

A further development of the second approach is a multiple-model-based molecular-replacement search (Buehler *et al.*, 2009). In this case, the output of the procedure is no longer the rigid-body parameters of the search model, as in traditional molecular replacement, but a phase set.

16.3.4.2. *Search targets – overview*

Two ways of generating phase variants lead to two different types of search targets. First, the phase selection can be applied at the level of the Fourier maps. For each trial phase set, a Fourier map can be calculated using experimental magnitudes. Then, the selection rule checks whether it has the features of a correct macromolecular Fourier synthesis or not. This approach has the advantage that it avoids model-related problems. If the phase sets are calculated from a model, then the corresponding calculated structure-factor magnitudes are also available. Their similarity to the experimental ones can be used to select the associated phase set.

16.3.4.3. *Histogram of a Fourier synthesis*

The histogram of a Fourier synthesis $\rho(\mathbf{r})$ indicates which values are present and how frequently these values appear in the synthesis (Lunin, 1988, 1993; Harrison, 1988; Luzzati *et al.*, 1988; Zhang & Main, 1990). To present the histogram numerically, the interval $(\rho_{\min}, \rho_{\max})$ of the possible values of $\rho(\mathbf{r})$ is divided into K equal parts (bins) and for every bin the frequency

$$\nu_k = n_k/N, \quad k = 1, \ldots, K, \tag{16.3.4.1}$$

is calculated. Here, N is the total number of grid points at which $\rho(\mathbf{r})$ is calculated and n_k is the number of grid points with $\rho(\mathbf{r})$ values belonging to the kth bin. The set of frequencies $\{\nu_k\}_{k=1}^{K}$ is called the histogram corresponding to the function $\rho(\mathbf{r})$. It depends on the resolution of the synthesis and is sensitive to phase errors. The 'standard histogram' $\{\nu_k^{\rm exact}\}_{k=1}^{K}$ at a particular resolution corresponds to the Fourier synthesis calculated with the observed magnitudes and the correct phases. It can be predicted before the phases are determined (Zhang & Main, 1990; Lunin & Skovoroda, 1991). When the standard histogram is known, it can be used to select appropriate Fourier syntheses. For example, a selection criterion may be defined as the correlation of the standard histogram with that from the Fourier synthesis calculated with the observed magnitudes and trial phases. Similar low-resolution phasing ideas have been used in protein electron crystallography (Dorset, 2000).

16.3.4.4. *Map connectivity*

The quality of phase sets may be also judged by the topological properties of regions of high electron density, *e.g.* their connectivity. A visual inspection of continuous regions in Fourier maps and the absence of noisy peaks were used for many years to estimate map quality. A formal scheme of the use of connectivity for phase improvement was discussed by Baker *et al.* (1993).

These ideas can be incorporated into low-resolution phasing as follows (Lunin, Lunina & Urzhumtsev, 2000). For a Fourier synthesis, $\rho(\mathbf{r})$, calculated with the observed magnitudes and trial phases, each trial phase set is associated with a mask region

$$\Omega_\kappa = \{\mathbf{r}\colon \rho(\mathbf{r}) > \kappa\} \tag{16.3.4.2}$$

composed of the grid points with the highest values of $\rho(\mathbf{r})$. The simplest characteristics of the region are the number of connected components in the region Ω_κ and their size. If the synthesis resolution is low and the cutoff level is high enough, it is expected that the region Ω_κ will consist of a small number of globs corresponding to isolated molecules. The correspondence between the connected components in the Ω_κ mask and the desired number (for example, a known number of molecules in the unit cell) may be used as a connectivity-based selection criterion. More sophisticated criteria can be introduced.

16.3.4.5. Few-atoms model approach

At low resolution, the content of a macromolecular crystal may be represented by a set of large isotropic Gaussian scatterers (pseudo-atoms). The number of scatterers necessary for appropriate modelling depends on the molecular shape and the resolution. Their size, represented by their isotropic displacement parameter (B factor), can be estimated at the first step of the procedure. The coordinates of the spheres can be generated randomly and uniformly in the unit cell. For an advanced search, when the molecular envelope is already known, the spheres can only be generated inside this envelope. More complicated generation rules can be applied as well. For a generated model, a set of structure factors $\mathbf{F}^{\rm calc}(\mathbf{s})$ is calculated and their magnitudes $F^{\rm calc}(\mathbf{s})$ are compared with the experimental data $F^{\rm obs}(\mathbf{s})$ (Lunin *et al.*, 1995). When these values are close enough, the corresponding phase set is selected for further analysis.

In the simplest case, when the molecule is approximated by a single sphere, a systematic search (instead of a random one) of the position of this sphere in the unit cell becomes possible (Jack *et al.*, 1975; Podjarny *et al.*, 1987; Harris, 1995; Andersson & Hovmöller, 1996). On the other hand, some attempts to obtain *ab initio* molecular envelopes when working with many pseudo-atoms have been also reported (Subbiah, 1991, 1993; David & Subbiah, 1994).

16.3.4.6. Likelihood-based selection

Statistical likelihood estimates the probability of reproducing experimental values within a framework of a statistical hypothesis. For example, the likelihood of a trial mask region may be estimated as the probability of obtaining the values of the observed magnitudes when placing atoms randomly and uniformly into the mask and calculating the structure factors from such a random model (Lunin *et al.*, 1998):

$$L = \text{Probability}\big\{\{F^{\rm calc}(\mathbf{s})\}\text{ are close to }\{F^{\rm obs}(\mathbf{s})\}\big\},\quad \mathbf{s}\in S_{\rm test}. \tag{16.3.4.3}$$

The most reasonable masks are those with the highest probability. In an advanced approach, more general prior probability distributions $p_{\rm prior}(\mathbf{r})$ for atomic coordinates are considered, and a search for the one that maximizes the probability is conducted [equation (16.3.4.3)] (Bricogne & Gilmore, 1990). Such an approach may be used to select a mask (or prior) from several alternatives.

The likelihood given by equation (16.3.4.3) becomes a selection criterion for phase sets if every trial phase set is associated with a mask or more general prior. A mask may be constructed as a region of highest values in a Fourier synthesis calculated with observed magnitudes and trial phases (Petrova *et al.*, 2000). A more advanced prior distribution may be associated with trial phases. For example, a search for the most featureless distribution (maximum entropy) among all the distributions that are consistent with the trial phases and observed magnitudes has been proposed (Bricogne & Gilmore, 1990). The likelihood given by equation (16.3.4.3) may be calculated in a straightforward Monte Carlo-type procedure (Petrova *et al.*, 2000) or with the use of an analytical approximation of the likelihood function (Lunin & Urzhumtsev, 1984; Bricogne & Gilmore, 1990). It should be noted that the set of reflections $S_{\rm test}$ used to calculate the likelihood given by equation (16.3.4.3) is generally different from the set of reflections $S_{\rm work}$ for trial phases.

16.3.4.7. Binary functions

Binary integer programming (BIP) is an effective approach for solving a system of linear inequalities in binary unknowns (0 or 1). Crystallographers are often interested in the study of a region with density values above a certain level, *i.e.* in a binary function representing this region.

The unknown phase values are linked to the observed magnitudes by nonlinear equations, but the equations may be reduced to linear ones if the phase values are restricted to a grid (Lunin, Urzhumtsev & Bockmayr, 2002). On the other hand, the use of binary representations of electron density and phases introduces additional errors in the calculations, reducing the advantages of the BIP approach. Another approach for obtaining binary masks at a grid uses genetic algorithms (Webster & Hilgenfeld, 2001).

16.3.4.8. Common features of selection criteria

All the targets reported above (as well as some others tried in order to select phase sets) show similar problems when applied at low resolution (Lunin, Lunina, Petrova *et al.*, 2000):

(i) The best value of a target may correspond to quite a poor phase set.
(ii) The target value for a good phase set may be significantly worse than the best target value.
(iii) Local refinement of model parameters or phases may lead to significant improvement of the target value without any improvement in the phases.

Owing to these problems, attempts to find the solution of the low-resolution phase problem by minimizing (or maximizing) a target are unreliable. At the same time, the selection of phase sets from a random ensemble by an appropriate target significantly increases the fraction of good ones.

16.3.5. Processing of the output

The output of the selection procedure may be presented as a two-dimensional table. In this table, each column represents one selected phase set and each row corresponds to one structure factor, represented by the phase value in different selected phase sets. If a selected phase set is considered as a point in multi-dimensional configuration space of all phase sets, the output becomes a ‘cloud’ in this space. A definition of the distance between two points in this space is required to study the detail of the distribution of these points (see Section 16.3.5.5).

16.3.5.1. Formal comparison of phase sets

Most frequently, two quantities are used to compare phase sets. One is the mean phase error,

$$D_\varphi = \frac{1}{M}\sum_{\mathbf{s}\in\mathbf{S}} \left|\varphi_{\mathbf{1}}(\mathbf{s}) - \varphi_2(\mathbf{s})\right|, \tag{16.3.5.1}$$

which takes into account the phase values only. The second quantity is the map correlation coefficient (Lunin & Woolfson, 1993),

$$\begin{aligned} C_\varphi &= \frac{\int\left[\rho_1(\mathbf{r}) - \langle\rho_1\rangle\right]\left[\rho_2(\mathbf{r}) - \langle\rho_2\rangle\right]\,{\rm d}V_{\mathbf{r}}}{\left\{\int\left[\rho_1(\mathbf{r}) - \langle\rho_1\rangle\right]^2\,{\rm d}V_{\mathbf{r}}\,\int\left[\rho_2(\mathbf{r}) - \langle\rho_2\rangle\right]^2\,{\rm d}V_{\mathbf{r}}\right\}^{1/2}} \\ &= \frac{\sum_{\mathbf{s}\in S} F^{\rm obs}(\mathbf{s})^2\cos\left[\varphi_1(\mathbf{s}) - \varphi_2(\mathbf{s})\right]}{\sum_{\mathbf{s}\in S} F^{\rm obs}(\mathbf{s})^2}, \end{aligned} \tag{16.3.5.2}$$

which weighs the differences by structure-factor phases. Here, M is the number of reflections in the studied set S, $\varphi_1(\mathbf{s})$ and $\varphi_2(\mathbf{s})$ are phase values corresponding to reflection $\mathbf{s}$ in two phase sets, $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ are Fourier syntheses calculated using the observed magnitudes and phases to be compared, and the sums in reciprocal space are calculated without the $F(\mathbf{0})$ term. Each of these quantities has its own disadvantage. The first is influenced by a number of weak reflections with poor phases that do not strongly affect the Fourier map but do significantly increase the value of D_φ. On the other hand, C_φ is influenced mostly by a small number of very strong reflections and is not so sensitive to errors in the phases of other structure factors. Sometimes, additional weighting in equations (16.3.5.1) and (16.3.5.2) or statistical calculations in resolution shells are applied to attenuate these effects.

If a collection of phase sets is considered as a cloud in multi-dimensional space, a formal distance may be defined on the basis of equation (16.3.5.2) as

$$\begin{aligned} d_\varphi &= \left\{ \frac{\int \left[\rho_1(\mathbf{r}) - \rho_2(\mathbf{r})\right]^2 \,\mathrm{d}V_\mathbf{r}}{\int \left[\rho_1(\mathbf{r})\right]^2 \,\mathrm{d}V_\mathbf{r}} \right\}^{1/2} \\ &= \left\{ \frac{\int \left[\rho_1(\mathbf{r}) - \rho_2(\mathbf{r})\right]^2 \,\mathrm{d}V_\mathbf{r}}{\int \left[\rho_2(\mathbf{r})\right]^2 \,\mathrm{d}V_\mathbf{r}} \right\}^{1/2} \\ &= \left[2\left(1 - C_\varphi\right)\right]^{1/2}, \end{aligned} \qquad (16.3.5.3)$$

varying from 0 to 2.

These formal measures of closeness are convenient when comparing phase sets with small differences but may be misleading at the initial stages of *ab initio* phasing, as discussed below.

16.3.5.2. *Phase ambiguity and alignment of phase sets*

Two randomly generated phase sets may result in Fourier maps $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ that differ only by an origin shift $\mathbf{t}$ permitted for the given space group: $\rho_2(\mathbf{r}) = \rho_1(\mathbf{r} - \mathbf{t})$. These two maps present the same object, they are calculated with the same experimental magnitudes and they are indistinguishable by the criteria discussed above. Therefore, they could both appear in the output of the selection procedure. However, the corresponding phases in these phase sets are different,

$$\varphi_2(\mathbf{s}) = \varphi_1(\mathbf{s}) + 2\pi(\mathbf{s}, \mathbf{t}), \qquad (16.3.5.4)$$

and an attempt to average such phase sets or superpose the corresponding maps is wrong. To compare and average these phase sets correctly they (or the corresponding maps) should be aligned (Lunin *et al.*, 1990; Hašek & Schenk, 1992; Lunin & Lunina, 1996). This alignment may be done on the basis of any formal measure of phase closeness, for example equations (16.3.5.1) or (16.3.5.2). It searches for the space-group-permitted shift of the second synthesis that results in the smallest phase difference between the two sets.

The enantiomer density transformation $\rho(\mathbf{r}) \to \rho(-\mathbf{r})$ conserves the features of the Fourier maps used in the selection criteria, conserves the magnitudes of the structure factors and substitutes the phases φ by $-\varphi$. Only high-resolution images might be able to solve the ambiguity and reveal the 'handedness' at later steps through the features of the secondary-structure elements. At the initial stage of phasing, the two enantiomer solutions may be considered equivalent and the enantiomer change $\varphi \to -\varphi$ may be added as an allowed operation in the alignment process.

Finally, a density transformation $\rho(\mathbf{r}) \to -\rho(\mathbf{r})$ preserves the magnitudes, changes the phases by π and flips the envelopes for a given cutoff level. In order to assess which of the two molecular envelopes is more probable, $\rho(\mathbf{r})$ or $-\rho(\mathbf{r})$, a generalized maximum-likelihood technique can be used (Lunin *et al.*, 1998).

The problem of phase alignment may be solved to some extent by fixing in advance the phase values for a small number (up to four, depending on the space group) of specially chosen reflections (reflections fixing the origin and/or the enantiomer). Nevertheless, for large phase sets the influence of these reflections may be too small to produce a reliable alignment.

In the following, when we discuss the closeness of two phase sets, we always suppose that they have already been aligned.

16.3.5.3. *Multiple alignment and 'central' phase set*

The problem of alignment becomes more complicated when applied to several phase sets simultaneously. The choice of a 'central' point of reference seems to be a reasonable solution, provided it minimizes the dispersion of the phase sets. To do so, let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}$ be a 'cloud' of selected phase sets considered as points in multidimensional configuration space. The point $\mathbf{v}^*$ is defined as the centre of the cloud if it minimizes the size of the cloud after all phase sets have been aligned to it:

$$Q(\mathbf{v}^*) = \sum_{n=1}^{N} d_\varphi^2(\mathbf{v}_n, \mathbf{v}^*) \Rightarrow \min, \qquad \mathbf{v}^* \in \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}, \qquad (16.3.5.5)$$

or

$$Q_1(\mathbf{v}^*) = \sum_{n=1}^{N} d_\varphi(\mathbf{v}_n, \mathbf{v}^*) \Rightarrow \min. \qquad (16.3.5.6)$$

This phase set may be used as the reference point to align the phase sets or as a current solution of the phase problem.

16.3.5.4. *Phase averaging and assignment of the probability distribution*

The simplest way to process the selected phase sets is to average them. For every reflection $\mathbf{s}$ from the set S, the best phase $\varphi^{\text{best}}(\mathbf{s})$ and its figure of merit $m(\mathbf{s})$ are calculated as

$$m(\mathbf{s}) \exp\left[i\varphi^{\text{best}}(\mathbf{s})\right] = \frac{1}{N} \sum_{n=1}^{N} \exp\left[i\varphi_n(\mathbf{s})\right], \qquad (16.3.5.7)$$

where $\varphi_n(\mathbf{s})$ is the phase in the nth selected set corresponding to the reflection $\mathbf{s}$, and N is the number of selected sets. The figure of merit

$$m(\mathbf{s}) = \frac{1}{N} \sum_{n=1}^{N} \cos\left[\varphi_n(\mathbf{s}) - \varphi^{\text{best}}(\mathbf{s})\right] \qquad (16.3.5.8)$$

reflects the divergence of the phase values of the reflection $\mathbf{s}$ in different selected phase sets.

The output of the phasing cycle may also be used to derive an approximate unimodal distribution,

$$\begin{aligned} P(\varphi) &= \frac{1}{I_0(T)} \exp[T \cos(\varphi - \theta)] \\ &= \Omega(A, B) \exp[A \cos\varphi + B \sin\varphi], \end{aligned} \qquad (16.3.5.9)$$

(also called 'von Mises', 'circular normal', 'Sim' *etc.*), with $\theta = \varphi^{\text{best}}$, $I_1(T)/I_0(T) = m$, or a bimodal distribution,

$$P(\varphi) = \Omega(A, B, C, D) \times \exp[A\cos\varphi + B\sin\varphi + C\cos 2\varphi + D\sin 2\varphi] \qquad (16.3.5.10)$$

(Hendrickson & Lattman, 1970). These distributions may be used to generate phases at the next cycle of phasing.

16.3.5.5. Cluster analysis

Accurate processing of the selected phase sets may involve methods of cluster analysis (Lunin *et al.*, 1990, 1995; Buehler *et al.*, 2009). Cluster analysis is a branch of applied mathematics aimed at classifying a set of points in a multidimensional space into several compact groups called clusters (or classes), so that the points inside a particular cluster are close to each other while different clusters are distant in space. Methods of cluster analysis use the matrix of point-to-point distances as input information. If the cluster analysis shows that the selected population can be split into several significantly different clusters, then the averaging is performed in each cluster separately, resulting in several alternatives for the solution of the phase problem and requiring multisolution strategies.

16.3.6. Conclusions and examples

The study of large complexes has renewed research interest in low-resolution phasing, since the corresponding crystals often diffract only to low resolution. *Ab initio* low-resolution phasing can give shape information in the unit cell and, in favourable cases, secondary-structure information. This information can be combined with other imaging techniques to obtain a reliable description of the structure.

The low-resolution maps for the ribosomal T50S particle obtained as early as the mid-1990s (Volkmann *et al.*, 1995; Urzhumtsev *et al.*, 1996; see also Lunin, Lunina *et al.*, 2002), for the low-density lipoprotein particle (Lunin *et al.*, 2001) and for some smaller macromolecules (Fokine *et al.*, 2003; Müller *et al.*, 2006) are some examples. In favourable cases, such images show not only the molecular envelope but also the secondary structure (Lunina *et al.*, 2003). At the same time, the current *ab initio* low-resolution phasing procedures are not yet completely automated and require significant human intervention.

References

Andersson, K. M. & Hovmöller, S. (1996). *Phasing proteins at low resolution. Acta Cryst.* D**52**, 1174–1180.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Uniqueness and the ab initio phase problem in macromolecular crystallography. Acta Cryst.* D**49**, 186–192.

Ban, N., Freeborn, B., Nissen, P., Penczek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. & Steitz, T. A. (1998). *A 9 Å resolution X-ray crystallographic map of the large ribosomal subunit. Cell*, **93**, 1105–1115.

Black, C. C. F., Fenn, R. H., North, A. C. T., Phillips, D. C. & Poljak, R. J. (1962). *Structure of lysozyme. A Fourier map of the electron density at 6 Å resolution obtained by X-ray diffraction. Nature (London)*, **196**, 1173–1178.

Bragg, W. L. & Perutz, M. F. (1952). *The external form of the haemoglobin molecule. Acta Cryst.* **5**, 277–283.

Bricogne, G. & Gilmore, C. J. (1990). *A multisolution method of phase determination by combined maximization of entropy and likelihood. I. Theory, algorithms and strategy. Acta Cryst.* A**46**, 284–297.

Buehler, A., Urzhumtseva, L., Lunin, V. Y. & Urzhumtsev, A. (2009). *Cluster analysis for phasing with molecular replacement: a feasibility study. Acta Cryst.* D**65**, 644–650.

Carter, C. W., Crumley, K. V., Coleman, D. E., Hage, F. & Bricogne, G. (1990). *Direct phase determination for the molecular envelope of tryptophanyl-tRNA synthetase from Bacillus stearothermophilus by X-ray contrast variation. Acta Cryst.* A**46**, 57–68.

Chapman, H. N., Fromme, P. & Spence, J. C. H. (2011). *Femtosecond X-ray protein nanocrystallography. Nature(London)*, **470**, 73–77.

Chapman, M. S., Tsao, J. & Rossmann, M. G. (1992). *Ab initio phase determination for spherical viruses: parameter determination for spherical-shell models. Acta Cryst.* A**48**, 301–312.

David, P. R. & Subbiah, S. (1994). *Low-resolution real-space envelopes: the application of the condensing-protocol approach to the ab initio macromolecular phase problem of a variety of examples. Acta Cryst.* D**50**, 132–138.

Dorset, D. L. (2000). *Low-resolution direct phase determination in protein electron crystallography – breaking globular constraints. Acta Cryst.* A**56**, 529–535.

Fokine, A., Morales, R., Contreras-Martel, C., Carpentier, P., Renault, F., Rochu, D. & Chabriere, E. (2003). *Direct phasing at low resolution of a protein copurified with human paraoxonase (PON1). Acta Cryst.* D**59**, 2083–2087.

Fourme, R., Shepard, W., Kahn, R., l'Hermite, G. & Li de La Sierra, I. (1995). *The multiwavelength anomalous solvent contrast (MASC) method in macromolecular crystallography. J. Synchrotron Rad.* **2**, 36–48.

Gilmore, C. J. (2000). *Direct methods and protein crystallography at low resolution. Acta Cryst.* D**56**, 1205–1214.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Guo, D. Y., Blessing, R. H. & Langs, D. A. (2000). *Globbic approximation in low-resolution direct-methods phasing. Acta Cryst.* D**56**, 1148–1155.

Hao, Q. (2006). *Macromolecular envelope determination and envelope-based phasing. Acta Cryst.* D**62**, 909–914.

Harris, G. W. (1995). *Fast ab initio calculation of solvent envelopes for protein structures. Acta Cryst.* D**51**, 695–702.

Harrison, R. W. (1988). *Histogram specification as a method of density modification. J. Appl. Cryst.* **21**, 949–952.

Harrison, S. C. & Jack, A. (1975). *Structure of tomato bushy stunt virus. Three-dimensional X-ray diffraction analysis at 16 Å resolution. J. Mol. Biol.* **97**, 173–191.

Hašek, J. & Schenk, H. (1992). *On the comparison of different sets of structure-factor phases. Acta Cryst.* A**48**, 693–695.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Hölzer, K., Weckert, E. & Schroer, K. (2000). *Properties of an electron-density map derived from a limited number of experimentally determined triplet phases. Acta Cryst.* D**56**, 322–327.

Jack, A., Harrison, S. C. & Crowther, R. A. (1975). *Structure of tomato bushy stunt virus. II. Comparison of results obtained by electron microscopy and X-ray diffraction. J. Mol. Biol.* **97**, 163–172.

Jamrog, D. C., Zhang, Y. & Phillips, G. N. Jr (2003). *SOMoRe: a multi-dimensional search and optimization approach to molecular replacement. Acta Cryst.* D**59**, 304–314.

Johnson, J. E., Akimoto, T., Suck, D., Rayment, I. & Rossmann, M. G. (1976). *The structure of southern bean mosaic virus at 22.5 resolution. Virology*, **75**, 394–400.

Kalinin, D. I. (1980). *Use of a cylindrical model of a protein to determine the spatial structure of the rhombic modification of leghaemoglobin. Sov. Phys. Crystallogr.* **25**, 307–313.

Lunin, V. Y. (1988). *Use of the information on electron density distribution in macromolecules. Acta Cryst.* A**44**, 144–150.

Lunin, V. Y. (1993). *Electron-density histograms and the phase problem. Acta Cryst.* D**49**, 90–99.

Lunin, V. Y. & Lunina, N. L. (1996). *The map correlation coefficient for optimally superposed maps. Acta Cryst.* A**52**, 365–368.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Skovoroda, T. P., Urzhumtsev, A. G. & Podjarny, A. D. (2000). *Low-resolution ab initio phasing: problems and advances. Acta Cryst.* D**56**, 1223–1232.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998). *On the ab initio solution of the phase problem for macromolecules at very low resolution. II. Generalized likelihood based approach to cluster discrimination. Acta Cryst.* D**54**, 726–734.

Lunin, V. Y, Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *On the ab initio solution of the phase* *problem for macromolecules at very low resolution: the few atoms model method. Acta Cryst.* D**51**, 896–903.

Lunin, V. Y., Lunina, N., Podjarny, A., Bockmayr, A. & Urzhumtsev, A. (2002). *Ab initio phasing starting from low resolution. Z. Kristallogr.* **217**, 668–685.

Lunin, V. Y., Lunina, N. L., Ritter, S., Frey, I., Berg, A., Diederichs, K., Podjarny, A. D., Urzhumtsev, A. & Baumstark, M. W. (2001). *Low-resolution data analysis for low-density lipoprotein particle. Acta Cryst.* D**57**, 108–121.

Lunin, V. Y., Lunina, N. L. & Urzhumtsev, A. G. (2000). *Connectivity properties of high-density regions and ab initio phasing at low resolution. Acta Cryst.* A**56**, 375–382.

Lunin, V. Y. & Skovoroda, T. P. (1991). *Frequency-restrained structure-factor refinement. I. Histogram simulation. Acta Cryst.* A**47**, 45–52.

Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Improvement of protein phases by coarse model modification. Acta Cryst.* A**40**, 269–277.

Lunin, V. Y., Urzhumtsev, A. & Bockmayr, A. (2002). *Direct phasing by binary integer programming. Acta Cryst.* A**58**, 283–291.

Lunin, V. Y., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Direct low-resolution phasing from electron-density histograms in protein crystallography. Acta Cryst.* A**46**, 540–544.

Lunin, V. Y. & Woolfson, M. M. (1993). *Mean phase error and the map-correlation coefficient. Acta Cryst.* D**49**, 530–533.

Lunina, N., Lunin, V. & Urzhumtsev, A. (2003). *Connectivity-based ab initio phasing: from low resolution to a secondary structure. Acta Cryst.* D**59**, 1702–1715.

Luzzati, V., Mariani, P. & Delacroix, H. (1988). *X-ray crystallography at macromolecular resolution: a solution of the phase problem. Makromol. Chem. Macromol. Symp.* **15**, 1–17.

Müller, J. J., Lunina, N. L., Urzhumtsev, A., Weckert, E., Heinemann, U. & Lunin, V. Y. (2006). *Low-resolution ab initio phasing of Sarcocystis muris lectin SML-2. Acta Cryst.* D**62**, 533–540.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960). *Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution obtained by X-ray analysis. Nature (London)*, **185**, 416–422.

Petrova, T. E., Lunin, V. Y. & Podjarny, A. D. (2000). *Ab initio low-resolution phasing in crystallography of macromolecules by maximization of likelihood. Acta Cryst.* D**56**, 1245–1252.

Podjarny, A. D., Rees, B., Thierry, J.-C., Cavarelli, J., Jesior, J. C., Roth, M., Lewitt-Bentley, A., Kahn, R., Lorber, B., Ebel, J.-P., Giegé, R. & Moras, D. (1987). *Yeast tRNAAsp–Aspartyl-tRNA synthetase complex: low resolution crystal structure. J. Biomol. Struct. Dyn.* **5**, 187–198.

Podjarny, A. D., Schevitz, R. W. & Sigler, P. B. (1981). *Phasing low-resolution macromolecular structure factors by matricial direct methods. Acta Cryst.* A**37**, 662–668.

Rayment, I., Baker, T. S., Caspar, D. L. & Murakami, W. T. (1982). *Polyoma virus capsid structure at 22.5 Å resolution. Nature (London)*, **295**, 110–115.

Roth, M., Lewit-Bentley, A. & Bentley, G. A. (1984). *Scaling and phase-difference determination in solvent contrast variation experiments. J. Appl. Cryst.* **17**, 77–84.

Sayre, D. (2008). *Report on a project on three-dimensional imaging of the biological cell by single-particle X-ray diffraction. Acta Cryst.* A**64**, 33–35.

Seibert, M. M., Ekeberg, T. & Hajdu, J. (2011). *Single mimivirus particles intercepted and imaged with an X-ray laser. Nature (London)*, **470**, 78–81.

Shepard, W., Kahn, R., Ramin, M. & Fourme, R. (2000). *Low-resolution phase information in multiple-wavelength anomalous solvent contrast variation experiments. Acta Cryst.* D**56**, 1288–1303.

Strop, P., Brzustowicz, M. R. & Brunger, A. T. (2007). *Ab initio molecular-replacement phasing for symmetric helical membrane proteins. Acta Cryst.* D**63**, 188–196.

Subbiah, S. (1991). *Low-resolution real-space envelopes: an approach to the ab initio macromolecular phase problem. Science*, **252**, 128–133.

Subbiah, S. (1993). *Low-resolution real-space envelopes: improvements to the condensing protocol approach and a new method to fix the sign of such envelopes. Acta Cryst.* D**49**, 108–119.

Urzhumtsev, A. & Podjarny, A. (1995). *On the solution of the molecular-replacement problem at very low resolution: application to large complexes. Acta Cryst.* D**51**, 888–895.

Urzhumtsev, A. G., Vernoslova, E. A. & Podjarny, A. D. (1996). *Approaches to very low resolution phasing of the ribosome 50S particle from Thermus thermophilus by the few-atoms-models and molecular-replacement methods. Acta Cryst.* D**52**, 1092–1097.

Valegård, K., Liljas, L., Fridborg, K. & Unge, T. (1991). *Structure determination of the bacteriophage MS2. Acta Cryst.* B**47**, 949–960.

Volkmann, N., Schlunzen, F., Urzhumtsev, A. G., Vernoslova, E. A., Podjarny, A. D., Roth, M., Pebay-Peyroula, E., Berkovitch-Yellin, Z., Zaytsev-Bashan, A. & Yonath, A. (1995). *On ab initio phasing of ribosomal particles at very low resolution. CCP4 Newslett.* **31**, 23–32.

Webster, G. & Hilgenfeld, R. (2001). *An evolutionary computational approach to the phase problem in macromolecular X-ray crystallography. Acta Cryst.* A**57**, 351–358.

Zhang, K. Y. J. & Main, P. (1990). *Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. Acta Cryst.* A**46**, 41–46.

International Tables for Crystallography (2012). Vol. F, Chapter 17.1, pp. 443–447.

PART 17. MODEL BUILDING AND COMPUTER GRAPHICS

Chapter 17.1. Macromolecular model building and validation using *Coot*

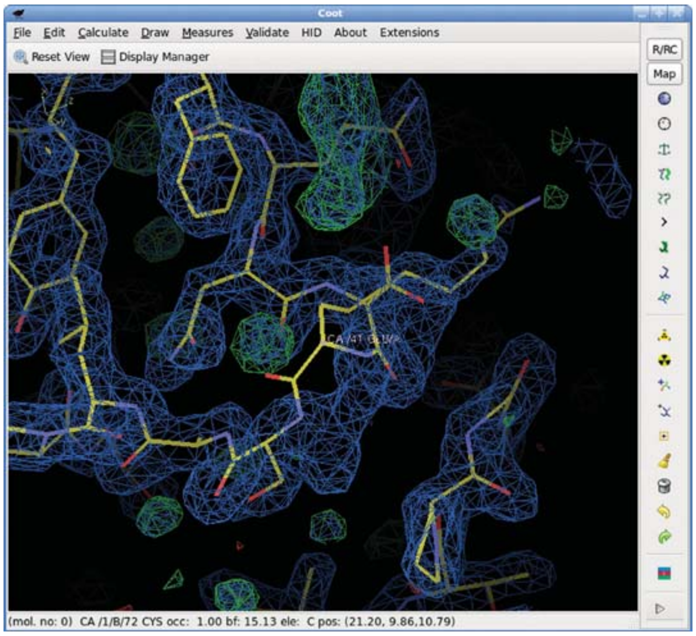
P. Emsley, B. Lohkamp and K. Cowtan

17.1.1. Introduction

Macromolecular model building using X-ray data is to some extent an interactive task, involving the iterative application of various optimization algorithms with the evaluation of the model and interpretation of the electron density by the scientist. *Coot* is an interactive three-dimensional molecular-modelling program particularly designed for building and validation of protein structures by facilitating the steps of the process.

In recent years, the initial construction of the protein chain has often been carried out using automatic model-building tools such as *ARP/wARP* (Langer *et al.*, 2008), *SOLVE/RESOLVE* (Wang *et al.*, 2004) and more recently *Buccaneer* (Cowtan, 2006). Therefore, relatively more time and emphasis has been placed on model validation than had previously been the case (Dauter, 2006). The refinement and validation steps become increasingly important and also more time consuming as data sets become poorer. *Coot* aims to provide convenient access to as many of the tools required in the iterative refinement and validation of a macromolecular structure as possible, to facilitate those aspects of the process which cannot be performed automatically. The software is also designed with a goal of being easy to learn, in order to provide a low barrier for scientists who are beginning to work with X-ray data (Fig. 17.1.1.1).

The principal tasks of the software are the visualization of macromolecular structures and data, the building of models into electron density, and the validation of existing models; these will be considered in the next three sections. The remaining sections of the chapter will deal with more technical aspects of the software, including interactions with external software, scripting and testing.



(mol. no: 0) CA /1/B/72 CYS occ: 1.00 bf: 15.13 ele: C pos: (21.20, 9.86,10.79)

Figure 17.1.1.1
The *Coot* main window. The main display area shows a molecule and electron density. At the top of the window is a menu bar providing access to most of the tools. Commonly used model-manipulation tools are also available through the toolbar on the right. Below the menu bar is an area for user-definable buttons.

17.1.2. Model building

Initial building of protein structures from experimental phasing is usually accomplished by automated methods. The main focus in *Coot*, therefore, is the completion of initial models generated by either molecular replacement or automated model building. However, the following features are provided for cases where an initial model is not available.

17.1.2.1. Tools for general model building

17.1.2.1.1. Cα baton mode

Baton building, introduced by Jones [see for example Kleywegt & Jones (1994)] allows a protein main chain to be built by using a 3.8 Å 'baton' to position successive α-carbons at the correct spacing. In *Coot*, this facility is reproduced and coupled with an electron-density ridge-trace skeleton (Greer, 1974). First, a skeleton is calculated following the ridges of the electron density. Then, the user selects baton-building mode, which places an initial baton with one end at the current screen centre. Candidate positions for the next α-carbon are highlighted as crosses, selected from those points on the skeleton which lie at the correct distance from the start point. The user can cycle through a list of candidate positions using the 'Try Another' button, or, alternatively, rotate the baton freely by use of the mouse. Additionally, the length of the baton can be changed to accommodate moderate inaccuracies in the α-carbon positions. Once a new position is accepted, the baton moves so that its base is on the new α-carbon. In this way, a chain may be traced manually at a rate of between 1 and 10 residues per minute.

Having placed the α-carbons, the rest of the main-chain atoms may be generated automatically using an implementation of the method of *CALPHA* (Esnouf, 1997).

17.1.2.1.2. Find secondary structure

Protein secondary-structure elements, including α-helices and β-strands, can be located by their repeating electron-density features, which lead to high and low electron-density values in characteristic positions relative to the consecutive α-carbons. The 'Find Secondary Structure' tool performs a six-dimensional rotational and translational search to find the likely positions of helical and strand elements within the electron density. This search has been highly optimized in order to achieve interactive performance for moderately sized structures, and, as a result, is less exhaustive than the corresponding tools employed in the automated model-building packages. However, it can provide a

Copyright © 2012 International Union of Crystallography

very rapid indication of map quality and a starting point for model building.

17.1.2.1.3. *Place helix here* and *Place strand here*

At low resolution it is sometimes possible to identify secondary-structure features in the electron density when the Cα positions are not obvious. If the user can identify that a helix or strand feature is present at a position in the electron-density map, *Coot* can place that feature automatically. This involves several stages. First, the orientation of the helix or strand is identified by using a rotational search to orient a cylindrical template to contain the maximum amount of density. Then, an ensemble of helix or strand fragments is placed in the density to determine the direction and registration of the helix or strand. The best fragment is extended to provide the longest element with a good match to the density.

17.1.2.1.4. *Find ligands*

The ability to fit ligands rapidly into electron-density maps is a frequently used technique, particularly useful to pharmaceutical companies (*e.g.* Williams *et al.*, 2005). The mechanism in *Coot* addresses a number of ligand-fitting scenarios and is a modified form of an algorithm described previously (Oldfield, 2001). It is common practice in 'fragment screening' to soak different ligand types into the same crystal (*e.g.* Blundell *et al.*, 2002). Using *Coot*, one can either specify a region in space or search a whole map for either a single or a number of different ligand types. In the 'whole map' scenario, candidate ligand sites are found by cluster analysis of a residual map and the candidate ligands are fitted in turn to each site, the candidate orientations being generated by matching the eigenvectors of the ligand to that of the cluster. Each candidate ligand is fitted and scored against the electron density. The best fitting orientation of the ligand candidates is chosen.

17.1.2.2. Rebuilding and refinement

The rebuilding and refinement tools are the primary means of model manipulation in *Coot*, and are all grouped together in the 'Model/Fit/Refine' tool set. These tools may be accessed either through a toolbar which is usually docked on the right-hand side of the main window, or through a separate 'Model/Fit/Refine' window containing buttons for each of the toolbar functions.

The core of the rebuilding and refinement tools is the real-space refinement (RSR) engine, which handles refinement of the atomic model against an electron-density map and regularization of the atomic model against geometric restraints. Refinement may be invoked both interactively when executed by the user, and non-interactively as part of some of the automated fitting tools. The refinement and regularization tools are supplemented by a range of additional tools aimed at assisting the fitting of protein chains. These features are discussed below.

17.1.2.3. Tools for moving existing atoms

17.1.2.3.1. Real-space refine zone

The real-space refine tool is the most frequently used tool in the refinement and rebuilding of atomic models, and is also incorporated as a final stage in a number of other tools, *e.g.* 'Add Terminal Residue . . . '. In interactive mode, the user selects the RSR button and then two atoms bounding a range of monomers (amino acids or otherwise). All atoms in the selected range of monomers will be refined, including any flanking residues. Atoms of the flanking residues are marked as 'fixed' but are required to be added to the refinement so that the geometry (*e.g.* peptide bonds, angles and planes) between fixed and moving parts is also optimized.

The selected atoms are refined to optimize a target consisting of two terms: the first being the Z-weighted sum of the electron-density values over all the atomic centres and the second being the stereochemical restraints on bond lengths, angles *etc*. The progress of the refinement is shown with a new set of atoms, displayed in white/pale colours. Once convergence is reached, the user is offered a dialogue box with a set of r.m.s. deviation scores and coloured 'traffic lights' indicating the current geometry scores in each of the geometrical criteria. Additionally, a warning is issued if the refined range contains new *cis*-peptide bonds.

The user may adjust the model by selecting an atom with the mouse and dragging it. The other atoms will move with the dragged atoms. Alternatively, a single atom may be dragged by holding the Ctrl key. As soon as the atoms are released, the model will refine from the dragged position. Additional atoms may be selected to be fixed during the refinement. This selection is made before starting the refinement.

17.1.2.3.2. Sphere refinement

One problem with the refinement mode described above was that it only considers a linear range of residues. This can cause problems, particularly at lower resolutions with some side chains inappropriately falling into electron density of neighbouring residues. Additionally, a linear residue selection precluded the refinement of entities such as disulfide bonds. Therefore, a new residue-selection mechanism was introduced to address these issues: the so called 'Sphere Refinement'. This mode selected residues that have atoms within a given radius of a specified position (typically within 4 Å of the centre of the screen). The residues are pairwise matched against the dictionary to find the appropriate restraint type.

17.1.2.3.3. Ramachandran restraints

At lower resolutions it is sometimes difficult to obtain an acceptable fit to the density and at the same time produce a Ramachandran plot of high quality (most residues in favourable regions and less than 1% outliers). If a Ramachandran score is added to the target function then the Ramachandran plot can be improved.

The analytical form for torsion gradients ($\partial\theta/\partial x_1$ and so on) for each of the x, y, z positions of four atoms contributing to the torsion angle has been reported previously (Emsley & Cowtan, 2004) (in the case of Ramachandran restraints, the θ torsions will be φ and ψ). The extension of the torsion gradients for use in Ramachandran restraints is performed using a two-dimensional log Ramachdran plot. This is generated as tables (one for each type of Pro, Gly and non-Pro or Gly), containing the log of the Ramachandran distribution. Where the Ramachandran probability goes to zero, the log probability would become infinite, and so is replaced by values which become increasingly negative with distance from the nearest nonzero value. This provides a weak gradient in the disallowed regions towards the nearest allowed region.

17.1.2.3.4. Rotamer tools

Four tools are available for the fitting of protein side chains. For a side chain whose amino-acid type is already correctly assigned, the best rotamer may be chosen to fit the density either automatically or manually. If the automatic option is chosen, then the side-chain rotamer from the *MolProbity* library (Lovell *et al.*,

2000) that gives rise to the highest density values at the atomic centres is selected and refined. Otherwise, the user is presented with a list of rotamers for that side-chain type, sorted by frequency in the database. Rotamers are named according to the *MolProbity* system (Lovell *et al.*, 2000).

The other two options, ‘Mutate & Auto Fit’ and ‘Simple Mutate’, allow the amino-acid type to be assigned or changed. The ‘Mutate & Auto Fit Rotamer’ option allows an amino-acid type to be selected from a list and then immediately performs the autofit rotamer operation as above. The ‘Simple Mutate’ option changes the amino-acid type and builds the side-chain atoms in the most frequently occurring rotamer, without refinement.

17.1.2.3.5. *Torsion editing*

Side-chain (or ligand) torsion angles need to be defined prior to editing. Either the user manually defines the four atoms forming the torsion angle (‘Torsion General’) or the torsion angles are determined from a dictionary description and the user selects the one to edit. In the latter case, the bond around which the selected torsion angle is edited is visually marked.

17.1.2.4. *Tools for adding atoms to the model*

17.1.2.4.1. *Find waters*

The water-finding mechanism in *Coot* uses the same mechanism for detecting unmodelled density as is used in the ligand fitting. However, only those features below a certain volume (by default 4.2 Å^3) are considered as candidate sites for water molecules. The centre of each feature is computed and then a distance check is made to the potential hydrogen-bond donors or acceptors in the protein molecule (or other waters). The distance criteria for an acceptable hydrogen-bond length is under user control. Additionally, a test for acceptable sphericity of the electron density is performed.

17.1.2.4.2. *Add terminal residue*

The *MolProbity* φ, ψ distribution is used to generate a set of randomly selected φ, ψ pairs. These random angles are used to generate the position of the N, Cα, O and C atoms of the next two residues. The conformation of these new atoms is then scored against the electron-density map and recorded. This procedure is carried out a number of times. The best-fitting conformation is offered as a candidate to the user (only the nearest of the two residues is kept).

17.1.2.5. *Tools for handling noncrystallographic symmetry (NCS)*

Noncrystallographic symmetry (NCS) can be exploited during the building of an atomic model, and also in the analysis of an existing model. In addition to some specific validation tools, *Coot* provides five tools to help with the building and visualization of NCS-related molecules:

(i) NCS ghost molecules. In order to visualize the similarities and differences between NCS-related molecules, a ‘ghost’ copy of any or all NCS-related chains may be superimposed over a chain in the model. The ‘ghost’ copies are displayed in thin lines and coloured differently as well as uniformly in order to distinguish them from the original. The superposition may be performed automatically by secondary-structure matching (Krissinel & Henrick, 2004), or by least-squares superposition.

(ii) NCS-averaged maps. In addition to viewing NCS-related copies of the electron density, the average density of the related regions may be computed and viewed. In noisy maps, this may provide a clearer starting point for model building.

(iii) NCS rebuilding. When building an atomic model of a molecule with NCS, it is often more convenient to work on one chain and then replicate the changes made to every NCS-related copy of that chain (at least in the early stages of model building).

(iv) NCS ‘jumping’: The view centre jumps to the next NCS-related peer chain and, at the same time, the NCS operators are taken into account so that the relative view remains the same. This provides a means for rapid visual comparison of NCS-related entities.

17.1.3. Validation

Coot incorporates a range of validation tools, including generic validation of the model against density, comprehensive geometrical validation tools for protein structures and additional validation tools specific to nucleotides. It also provides convenient interfaces to external validation tools, most notably the *MolProbity* suite (Davis *et al.*, 2007; Chapter 21.6), but also to the *Refmac* refinement software (Murshudov *et al.*, 1997) and dictionary (Vagin *et al.*, 2004).

Many of the internal validation tools provide a uniform interface in the form of colour-coded bar charts, for example the ‘Density fit analysis’ chart (Fig. 17.1.3.1). This window contains one bar chart for each chain in the structure. Each chart contains one bar for each residue in the chain. The height and colour of the bar both indicate the ‘goodness’ of the monomer, with small green bars being good or expected/conventional and large red bars being bad or unconventional. The chart is active, *i.e.* on mousing over the bar, tooltips provide relevant statistics, and clicking on a bar changes the view in the main graphics window to centre on the selected residue. In this way, a rapid overview of model quality is obtained and problem areas can be easily accessed. In order to obtain a good structure for submission, the user may simply cycle though the validation options, correcting any problems found.

The available validation tools are described in more detail in the following sections.

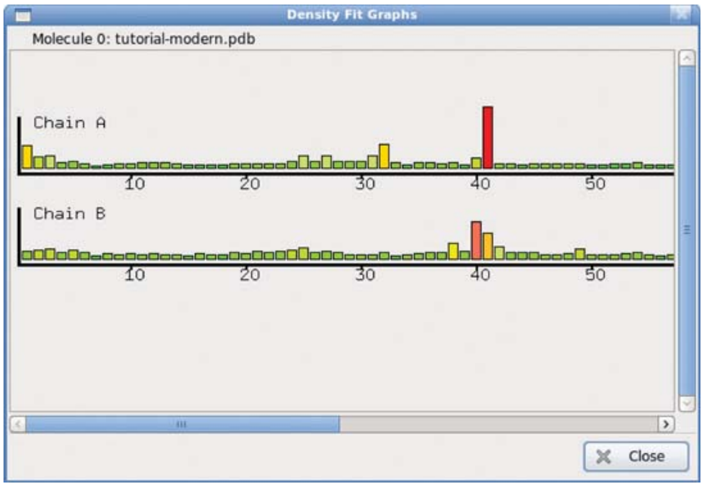


Figure 17.1.3.1
A typical validation graph. Bars represent individual residues in a chain, with an indication of quality for that residue given by both size and colour. The plot is interactive, *i.e.* clicking on a bar takes the user to the corresponding residue in the main window.

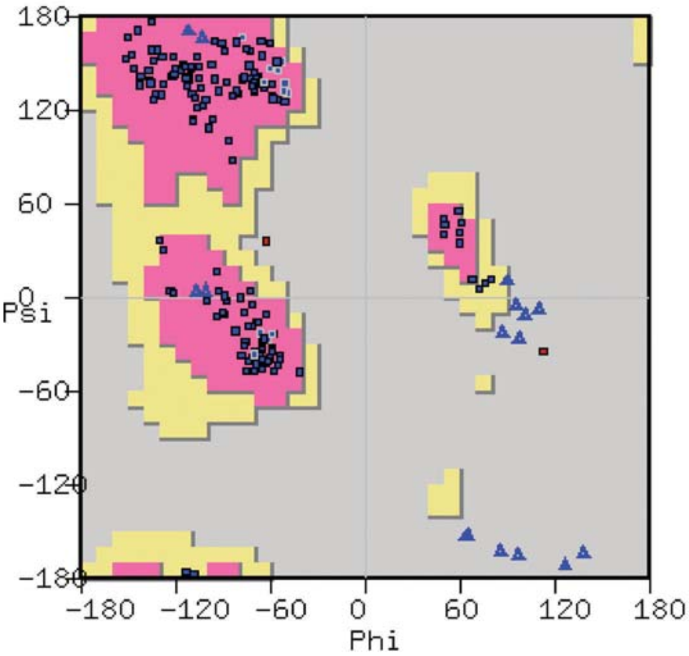


Figure 17.1.3.2
Interactive Ramachandran plot. The axes show the φ and ψ angles. Preferred regions are coloured in pink, allowed regions in yellow, and the background in grey for disallowed regions. Standard residues are shown as dark-blue squares, Pro residues as light-blue squares and Gly residues as light-blue open triangles. Residues in the disallowed regions are coloured red. Below the graph, in the *Coot* dialogue box, a summary of the number (and percentages) of residues in the different regions of the Ramachandran plot is given.

17.1.3.1. Ramachandran plot

The Ramachandran plot tool (Fig. 17.1.3.2) launches a new window in which the Ramachandran plot for the active molecule is displayed. A data point appears in this plot for each residue in the protein, with different symbols distinguishing Gly and Pro residues. The background of the plot shows frequency data for Ramachandran angles using the Richardsons' data (Lovell *et al.*, 2003).

The plot is interactive: clicking on a data point moves the view in the three-dimensional canvas to centre on the corresponding residue. Similarly, selecting an atom in the model highlights the corresponding data point. Moving the mouse over a data point corresponding to a Gly or Pro residue causes the Ramachandran frequency data for that residue type to be displayed.

17.1.3.2. Geometry analysis

The geometry (bonds, angles, planes) for each residue in the selected molecule is compared to dictionary values (typically provided by the mmCIF *Refmac* dictionary). Torsion angles are not analysed (there are other validation tools for torsion angles).

The statistics for the geometry graph are the average Z value for each of the geometry terms for that residue (peptide geometry distortion is shared between neighbouring residues). The tooltip on the geometry graph describes the geometry features giving rise to the highest Z value.

17.1.3.3. Peptide ω analysis

This validation tool for peptide ω torsion angles produces a graph marking the deviation from 180 degrees of the peptide ω angle. The deviation is assigned to the residue to include the C and O atoms of the peptide link. Thus, peptide ω angles of 90° are very bad. Optionally, ω angles of 0° can be considered ideal (for the case of intentional *cis*-peptide bonds).

17.1.3.4. Rotamer analysis

The rotamer statistics are generated from the nearest conformation in the *MolProbity* rotamer probability distribution (Lovell *et al.*, 2000). The height of the bar is inversely proportional to the rotamer probability.

17.1.3.5. Density-fit analysis

The bars in the density-fit graphs are inversely proportional to the average Z-weighted electron density at the atom centres and to the grid sampling of the map (*i.e.* maps with coarser grid sampling will have lower bars than a more finely gridded map, all other things being equal). Accounting for the grid sampling allows lower-resolution maps to have a density-fit graph without many or most residues being marked as worrisome owing to their atoms being in generally low levels of density.

17.1.4. Scripting

Many internal functions in *Coot* are accessible *via* an automically generated (SWIG) interface to the scripting languages Python (http://www.python.org) and Guile (a scheme interpreter; Kelsey *et al.*, 1998; http://www.gnu.org/software/guile/guile.html). Some of *Coot*'s graphics widgets are available for scripting *via* the same interface.

17.1.5. Discussion

Coot combines a range of modern methods in macromolecular model building and validation with a modern graphical user interface, following current design conventions to achieve ease of use, productivity, aesthetics and forgiveness. This is an ongoing process and although improvements can still be made, we believe that *Coot* provides a shallow learning curve combined with a high level of crystallographic awareness to aid the novice and non-specialist user, while providing a powerful and configurable tool set for the expert.

References

Blundell, T. L., Jhoti, H. & Abell, C. (2002). *High-throughput crystallography for lead discovery in drug design*. *Nat. Rev. Drug Discov.* **1**, 45–54.

Cowtan, K. (2006). *The Buccaneer software for automated model building. 1. Tracing protein chains*. *Acta Cryst.* D**62**, 1002–1011.

Dauter, Z. (2006). *Current state and prospects of macromolecular crystallography*. *Acta Cryst.* D**62**, 1–11.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids*. *Nucleic Acids Res.* **35**, W375–W383.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics*. *Acta Cryst.* D**60**, 2126–2132.

Esnouf, R. M. (1997). *Polyalanine reconstruction from C_α positions using the program CALPHA can aid initial phasing of data by molecular replacement procedures*. *Acta Cryst.* D**53**, 665–672.

Greer, J. (1974). *Three-dimensional pattern recognition: An approach to automated interpretation of electron density maps of proteins*. *J. Mol. Biol.* **82**, 279–301.

Kelsey, R., Clinger, W. & Rees, J. (1998). *Higher-Order Symb. Comput.* **11**, 7–105.

Kleywegt, G. J. & Jones, T. A. (1994). In *From First Map to Final Model. Proceedings of the CCP4 Study Weekend*, edited by S. Bailey, R. Hubbard & D. Waller, pp. 59–66. SERC Daresbury Laboratory.

Krissinel, E. & Henrick, K. (2004). *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions*. *Acta Cryst.* D**60**, 2256–2268.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7*. *Nat. Protoc.* **3**, 1171–1179.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Structure validation by Cα geometry: φ, ψ and Cβ deviation*. *Proteins*, **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *The penultimate rotamer library*. *Proteins*, **40**, 389–408.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method*. *Acta Cryst.* D**53**, 240–255.

Oldfield, T. J. (2001). *X-LIGAND: an application for the automated addition of flexible ligands into electron density*. *Acta Cryst.* D**57**, 696–705.

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use*. *Acta Cryst.* D**60**, 2184–2195.

Wang, J. W., Chen, J. R., Gu, Y. X., Zheng, C. D., Jiang, F., Fan, H. F., Terwilliger, T. C. & Hao, Q. (2004). *SAD phasing by combination of direct methods with the SOLVE/RESOLVE procedure*. *Acta Cryst.* D**60**, 1244–1253.

Williams, S. P., Kuyper, L. F. & Pearce, K. H. (2005). *Recent applications of protein crystallography and structure-guided drug design*. *Curr. Opin. Chem. Biol.* **9**, 371–380.

International Tables for Crystallography (2012). Vol. F, Chapter 17.2, pp. 448–458.

Chapter 17.2. Molecular graphics and animation

A. J. Olson

17.2.1. Introduction

Visualizing the unseeable world of molecules is the fundamental goal of crystallographic structure determination. Thus there is a natural synergy between the science of unravelling molecular structure and the technology of representing it. Graphics have always had a significant role in the analysis of diffraction data, the synthesis of molecular models and the communication of the information and knowledge gained in these scientific pursuits. At least since the time of René Häuy in the 18th century, crystallographers have attempted to use graphics and physical models to understand and explain the underlying nature of the solid state (Fig. 17.2.1.1). Over the years, crystallography has pushed the development of new technologies to aid in structure solution, and crystallographers have been early adaptors of new technologies. No technology has had more impact on crystallography than electronic computing. Nowhere is that impact more apparent than in what we have been able to study and how we have been able to visualize our structural results. With the pervasiveness of three-dimensional computer graphics in many aspects of everyday life, it is easy to forget the role that X-ray crystallography has played in its genesis and the role that graphics technology continues to play in the advancement of molecular structure analysis.

The human genome project and other efforts in biology and medicine have produced heightened emphasis on molecular depictions of increasing complexity. Visualization of such systems through computer-graphics technology is a key component in our understanding of these data and the models that we use to explain them. In fact, modelling and visualization techniques provide a bridge between experimental data at different scales, enabling placement of detailed atomic models of molecules from crystallography into lower-resolution data on large assemblies from electron microscopy or scanning probe imaging.

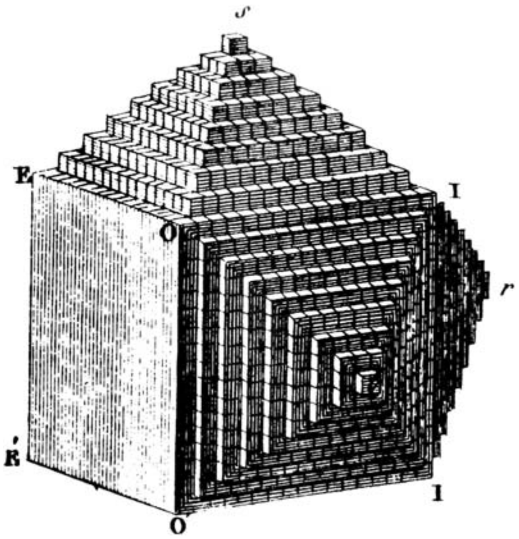


Figure 17.2.1.1
Model of a crystal structure proposed by René Häuy in *Traite elementaire de Physique*, Vol. 1 [Paris: De L'Imprimerie de Delance et Lesueur, 1803]. This model, proposed in 1784, was the first to connect the external facets of a crystal with an underlying regular arrangement of building blocks.

17.2.2. Background – the evolution of molecular graphics hardware and software

The complexity of molecular structure and the fact that these sub-microscopic objects of study are not directly visible have necessitated the use of physical or pictorial representations to aid in interpretation, manipulation and understanding. Illustrations and models made of wood, plastic or metal served these purposes from the development of the original theories of molecular structure through to the first nucleic acid and protein structures solved in the 1950s. Over the following years, computer graphics has evolved into a significant and ubiquitous technology, helping to sustain the explosive growth of macromolecular structure research. Today, computer graphics pervade the activities of much molecule-based research, from quantum chemistry to molecular biology.

Computer-based molecular graphics can be traced back to 1948 and the X-RAC project of R. Pepinsky at Pennsylvania State University (Pepinsky, 1952). Pepinsky developed an analogue computer to carry out the Fourier transformation of X-ray structure factors to produce electron-density maps. Integrated within X-RAC was an oscilloscope that could display the contours of the electron density (Fig. 17.2.2.1). These displays were, to my knowledge, the first computer-generated images of molecular structure. Crystallographers from around the world came to Pennsylvania State University to use X-RAC and to marvel at the speed and automation possible in the solution of molecular structures. While the digital revolution quickly overtook the analogue approach, X-RAC clearly set the precedent for molecular scientists as early implementors and adaptors of computational and graphics technology.

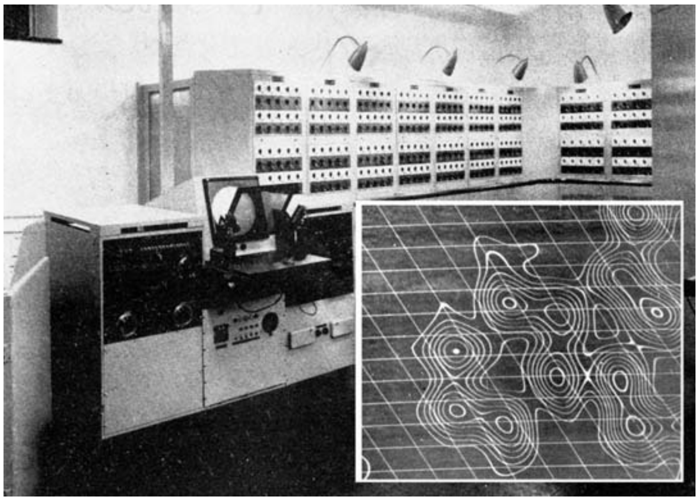


Figure 17.2.2.1
One bay of X-RAC showing coefficient panels and the display oscilloscope. Inset: photo from the oscilloscope, showing a region of the phthalocyanine Fourier map. Reproduced from Pepinsky (1952).

Copyright © 2012 International Union of Crystallography

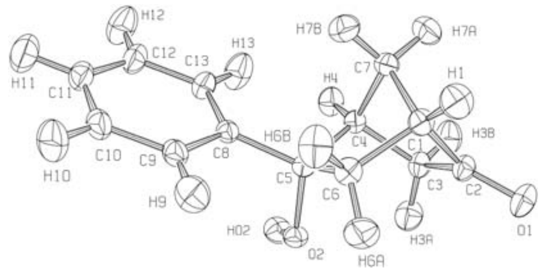


Figure 17.2.3.1
ORTEP plot of phenylhydroxynorbornanone showing atomic thermal ellipsoids (from *Thermal ellipsoid analysis: the fossil footprints of restless atoms*, by Carroll K. Johnson and Michael N. Burnett, Buerger Award Lecture at the ACA meeting in St. Louis, July 20–26, 1997).

In the 1960s, two seminal projects laid the foundation for modern molecular graphics. Early in the decade, Johnson's *ORTEP* (Johnson, 1970) program became widely available, allowing crystallographers to produce illustrations of three-dimensional (3-D) molecular structures on a pen plotter. These black-and-white line drawings of ball-and-stick models were used both for working drawings during structure analysis and for creating illustrations for publication. A few years later, experiments lead by Levinthal under Project MAC (Levinthal, 1966) at MIT pioneered the interactive display and transformation of 3-D molecular structures on a computer screen. By the end of the decade, the groundwork for molecular graphics was set: *ORTEP* convincingly demonstrated to a large number of scientists that the computer could be used as an alternative to the human hand to produce accurate drawings and stereoscopic pairs for the analysis and communication of the results of structural research. Project MAC showed that the computer could be used as an interactive environment in which to model and simulate on the molecular scale. These two projects helped define the two broad functions of molecular graphics: publication graphics, for which clarity of presentation is the essential goal, and working graphics, for which rapid feedback and high interactivity are the key elements.

In the 1970s, 3-D interactive computer-graphics systems became commercially available. Hardware offerings from companies such as Evans and Sutherland, Vector General, and Adage prompted a number of laboratories to develop interactive molecular-modelling software. Several of these early systems were devoted to the task of building an atomic model of a protein into the crystallographically derived electron-density map. Programs such as *BILDER* (Diamond, 1982), *MMSX* (Barry & McAlister, 1982), *FRODO* (Jones, 1978) and *GRIP* (Wright, 1982) began to replace metal Kendrew models and the cumbersome 'Richards Box' optical comparitors. This application, more than any other, sold these expensive (>$100 000) monochrome line-drawing graphics devices to the molecular-research community. Moreover, during that time, biomolecular structure determination was a major civilian consumer of 3-D interactive graphics devices.

Technical, commercial and scientific advances in the 1980s prompted enormous growth in the use of molecular graphics. As late as 1983, a worldwide list of laboratories using high performance graphics computers for molecular work could be maintained – the number was below 100 (Olson, 1983). By the end of the decade, that number grew into the thousands, and utilization spread beyond any ability to track it. At the beginning of the decade the expensive vector-graphics terminals were the only way to achieve interactive 3-D display. Ten years later, the colour raster display had taken over the interactive computer-graphics market, driving prices down and broadening display capabilities from lines and dots to include shaded surface representation. Early in the 1980s, several academic software packages, such as *GRAMPS/GRANNY* (O'Donnell & Olson, 1981; Connolly & Olson, 1985), *MIDAS* (Ferrin *et al.*, 1988) and *HYDRA* (Hubbard, 1986), went beyond electron-density fitting to provide general graphics functionality for examining molecular structure and properties. Over the decade, the remarkable evolution of computer hardware – the advent of microprocessors, very large scale integration (VLSI) devices, personal computers and scientific workstations – increased the accessibility of molecular graphics. By the mid-1980s, the demand was such that several commercial companies had been established to market molecular graphics and modelling software. By the end of the decade, structural scientists in academic and industrial research settings had a wide variety of use-tested hardware and software platforms with which to perform molecular modelling.

The 1990s witnessed remarkable advances in the technology and sociology of computing as well as in the science of molecular structure and design. Moore's law of the microcosm, which estimates that 'the effectiveness of microprocessors doubles every 18 months', continued to track growth accurately. Thus the performance-to-cost ratio of late 1990s computers was a millionfold higher than those of the mid-1960s. A 1997 200 Nintendo-64 game machine was faster and had more memory and far superior graphics than the Control Data 6600 supercomputer and peripherals of the 1960s, and, with its optional ($70) disk, bettered almost every technical specification of a 1980 VAX 11/780. Gelder's law of the telecosm posited in 1993 that 'bandwidth will treble every year for at least the next 25 years'. This, coupled with Metcalf's law, that 'the total value of a network to its users grows as the square of the total number of users' implies that the 'teleputer', or non-localized computing, is becoming the computational environment of the future.

While software development continues to lag behind hardware growth, the emergence of the World Wide Web and the concepts of network-based computing have catalysed a rethinking of the nature of software, its development, distribution and inter-operability. The concepts of cyberspace and 'virtual reality' have been implanted into the minds and expectations of the general public, promoting a renaissance in user-interface exploration and development. It is transforming the computer from a window through which to look into a portal through which to step. Suddenly, the other senses – sound, touch, taste and smell – can become part of the computational experience.

17.2.3. Representation and visualization of molecular data and models

In the early days of molecular computer graphics, the major goal was to represent the spatial structures of molecules, principally the locations of the atom centres and the covalent connectivity between them. Using X-ray diffraction analysis, one would first plot the electron density as line contours projected onto a plane and locate the atom centres from multiple projections. The molecule would then be represented by a simple bond diagram. As experimental and computational methods advanced, other representations were used to convey additional information about the structure. Johnson's *ORTEP* program plotted the thermal ellipsoids of each of the atoms, visualizing the magnitude and direction of their thermal vibrations as derived from the anisotropic temperature factors (Fig. 17.2.3.1). As colour raster

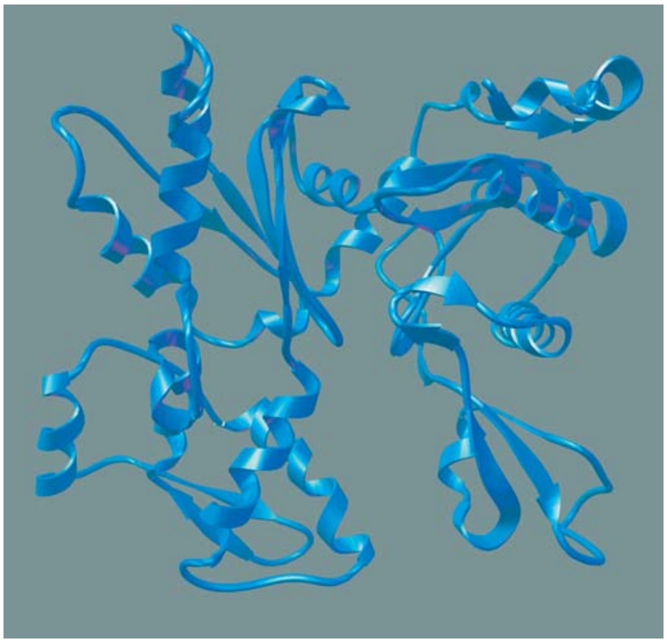


Figure 17.2.3.2
Ribbon diagram of actin monomer (PDB code: 1atn) (Kabsch *et al.*, 1990).

displays became available, space-filling CPK representations were used to visualize molecular shape and volume while using an atom-based colour scheme to show atomic composition and distribution (Porter, 1979) (see Fig. 17.2.3.4). The complexity of protein molecules prompted the introduction of simplified representations that replaced the all-atom visualization with tubes or ribbons (Branden *et al.*, 1975; Carson, 1991) (Fig. 17.2.3.2) that represented the fold of the protein chain. This simplification allowed the comparison of protein folds, and led to the beautiful classification of protein motifs by Richardson (1981).

As more structural information has become available, and as computational hardware technology has advanced, the ability to visualize a variety of molecular properties has become possible. Meanwhile, issues of interactivity, intelligibility and interpretability have become increasingly important as the systems under study have become more complex. There are three general approaches to visualizing the structures, properties and relationships of molecular systems: geometric construction, direct volumetric rendering and generic information visualization. Today almost all macromolecular modelling and visualization work is done using geometric representations of bonds, ribbons and surfaces, which are annotated by colour to represent atom type, chain characteristics, or electrostatic potential. For these purposes, there are several 'turn-key' programs that facilitate display and interaction. Programs such as *RASMOL* (Sayle & Milner-White, 1995), *GRASP* (Nichols *et al.*, 1995) and *MOLSCRIPT* (Kraulis, 1991) are widely used by the molecular-structure community. Some of the fundamentals of representation used in these types of programs, as well as more exploratory techniques, are described below.

While the world is continuous, our measurements of it tend to be finite and sampled. Thus data are usually represented as discrete values on a line, plane, volume or hypervolume. On the other hand, in order to capture the nature of the world, our models tend to be represented as continuous functions. Geometric construction is useful for rendering continuous models, while other techniques, such as volume rendering, lend themselves to the visualization of discrete data.

17.2.3.1. *Geometric representation*

Geometric construction encompasses dots, lines and surfaces described by lists of three-dimensional coordinates and connectivity or by analytic or parametric expressions that can generate such information for rendering. Basically, geometric rendering involves a projection of the 3-D geometry onto a two-dimensional viewing plane using matrix transformations that account for the viewpoint, perspective and clipping within the viewing volume. For dots and lines, the computation may end there; the only depth information in the rendering might be geometric perspective. Additional depth information can be added by 'atmospheric perspective' or depth cueing, where the brightness or colour is modulated by the depth values of the points (Fig. 17.2.3.3). Surface representations permit additional three-dimensional cues such as occlusion and shape-from-shading. Occlusion, or 'hidden surface removal', and atmospheric perspective depend on maintaining depth information for all of the picture elements (pixels) in screen space. Such 'depth-buffer' algorithms provide visibility information for a given viewpoint. Hardware *z*-buffers facilitate such calculations in the graphics pipeline. Lighting cues, such as shading, are attained by approximating the ambient, diffuse and specular reflectance of

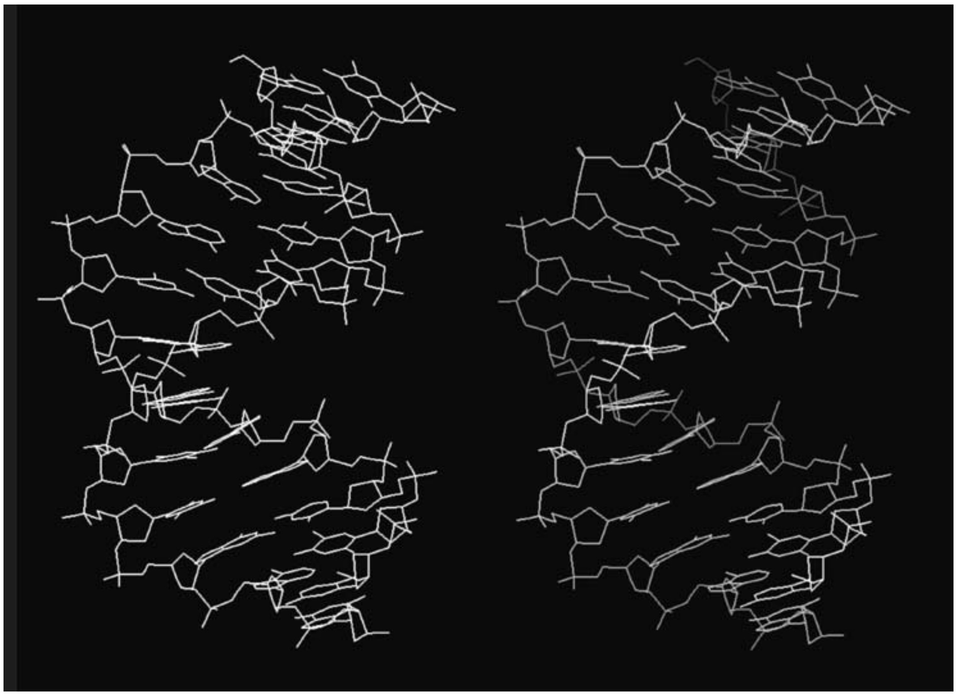


Figure 17.2.3.3
Simple bonding diagram of a DNA structure (PDB code: 140D) (Mujeeb *et al.*, 1993). On the left all lines are of equal intensity. On the right the lines are depth-cued to show which parts of the structure are closer to the viewer.

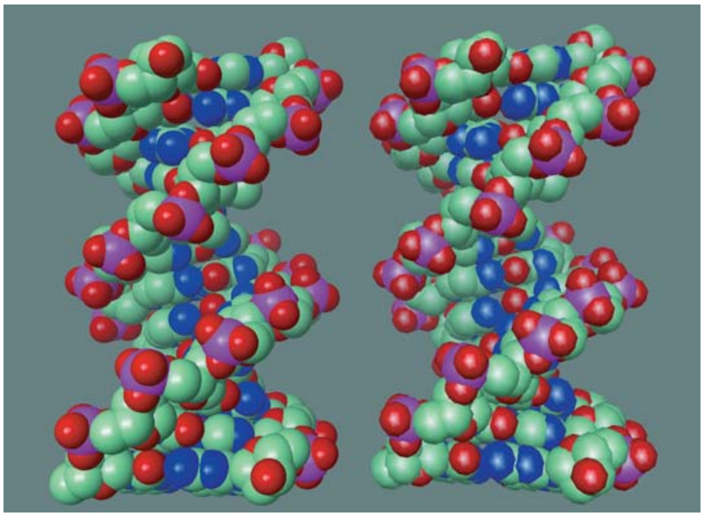


Figure 17.2.3.4
CPK representation of the same DNA structure as in Fig. 17.2.3.3. The model on the left uses highly tessellated spheres for the atoms, while the one on the right uses a coarser tessellation. The Gouraud shading model produces some lighting artifacts, such as the star-shaped highlights, which are most apparent on the right-hand figure. This is due to colour interpolation between the facet vertices.

the geometry using Lambert's law. Because typical surfaces are composed of polyhedral facets, interpolation schemes are used to produce smooth shaded representations. The most common technique used for molecular graphics is known as Gouraud shading (Gouraud, 1971), which interpolates the shaded colour values assigned at the vertices across the polyhedral face (Fig. 17.2.3.4). Phong shading (Phong, 1975), a more accurate but costly technique, interpolates the values of the normals of the facets to produce a more realistic rendering. Shading templates for specific geometries, such as spheres, can give very smooth results without having to resort to large polyhedral descriptions for each sphere. In the past, this approach was implemented in the graphics hardware design, resulting in very fast sphere rendering for molecular applications. With the advent of consumer-level 3-D graphics these specialized features have become increasingly rare. Shadows may also provide useful three-dimensional cues in viewing molecular objects, but may also be confusing when they provide too much visual contrast or clutter. Ray tracing is a general technique for producing a complete reflectance and shadow rendering of a three-dimensional scene. It can, however, be very costly in computational time, since every light ray in the final image must be iteratively traced back to its source. Faster approximations for shadow rendering have been implemented that work well for molecular scenes (Gwilliam & Max 1989; Lauher, 1990).

A number of useful surface representations have been developed that describe the interaction of a molecule with the surrounding solvent. Perhaps the most widely used are the solvent-accessible surface (Lee & Richards, 1971) and the molecular surface (Richards, 1977; Connolly, 1983; Sanner *et al.*, 1996), sometimes referred to as the Connolly or solvent-excluded surface (Fig. 17.2.3.5). For large molecules, such as proteins, which have many atoms buried from solvent, these surfaces have proven to be important in studying molecular interactions. They not only help to visualize the complementarity of interacting molecules, but they are also important in quantifying the entropic changes associated with solvent effects upon binding.

Surface representations have opened up the possibilities of displaying a large variety of computed or experimental molecular properties by mappings onto the surface using colour coding. Electrostatic potential, hydrophobicity, sequence conservation, surface shape and any other characteristic of the molecule that can be projected onto the surface can be colour coded and displayed. Typically, this is accomplished by colouring the vertices of the surface mesh using a colour mapping or scale and interpolating the colour across the polygonal faces of the mesh. Since colour values are interpolated between vertices, this can produce unwanted colour artifacts if there are abrupt spatial changes in the properties displayed, or if the colour interpolation does not correspond to the property mapping (Fig. 17.2.3.6).

Another method for projecting information on a surface is texture mapping, an approach that is analogous to applying an image 'decal' onto the surface. In this approach, instead of assigning colours to the surface vertices, indices are assigned which serve as coordinates into the image to be mapped. Thus, a great amount of detail may be displayed on a surface mesh that has relatively few polygons describing the geometry. Texture mapping has been used extensively in highly interactive graphics, such as flight simulators and video games, since transformation of the geometry tends to be the computational bottleneck. Since texture mapping requires an indexing scheme that relates an image to a set of geometric vertices on the molecular surface, one needs a

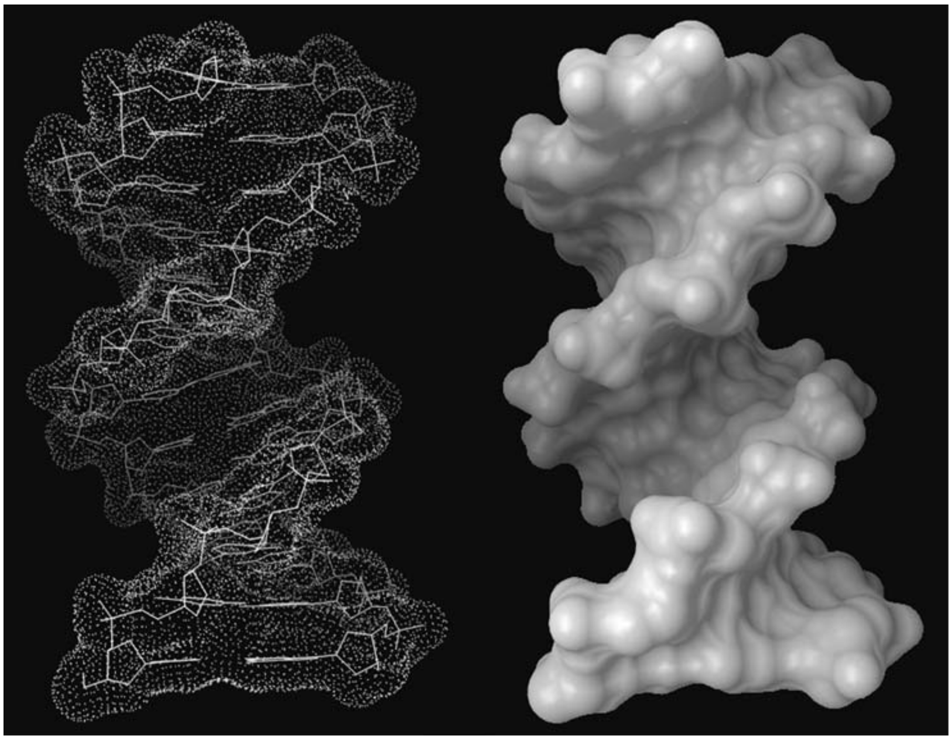


Figure 17.2.3.5
Solvent-excluded surface of the DNA structure using a water probe radius of 1.5 Å. The figure on the left shows a depth-cued dot surface, while the figure on the right shows a Gouraud-shaded triangulated surface.

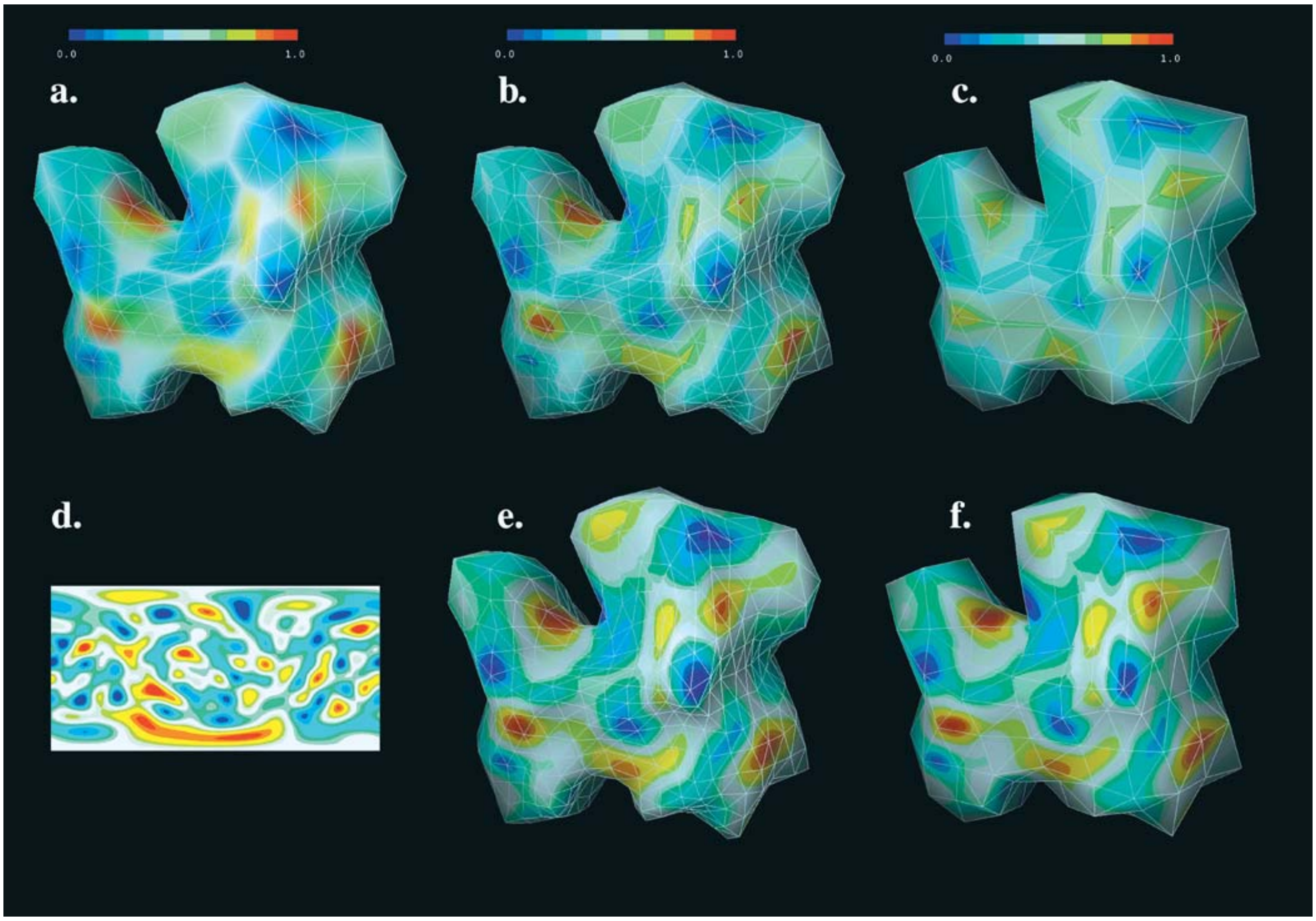


Figure 17.2.3.6

Hydrophobicity mapped onto a molecular surface. A spherical-harmonic approximation of the actin monomer solvent-excluded surface is shown. (*a*) Vertex colouring of a medium-mesh tessellated surface. The hydrophobicity colour scale is shown above. Notice that the colours blend between vertices, producing colour artifacts in relationship to the property scale. (*b*) The same medium-mesh representation as in (*a*) but using a property-based (one-dimensional) texture map, applying the same colour scale. Notice that the boundaries between the colours are distinct, even when they intersect vertices. Here the property value is interpolated. (*c*) A coarser mesh showing the same texture-mapping technique used in (*b*). Since the properties are only sampled at the vertices of the mesh, the finer details of the mapping are lost at this coarse triangulation. (*d*) A two-dimensional texture map created as a 'Mercator-like' projection in spherical coordinates (θ, φ) from the same hydrophobicity scale used in (*a*)–(*c*). (*e*) The 2-D texture map shown in (*d*) mapped onto the medium-mesh actin surface. Notice that the linear nature of the interpolation seen in (*b*) using the same mesh is no longer present. (*f*) The same 2-D texture map applied to the coarse-mesh surface of actin. Notice that, unlike in (*c*), the detail of the texture map is preserved independent of the mesh.

rational way of producing such a map. For one-dimensional texture maps, this is relatively easily accomplished by assigning the texture index of each vertex to an appropriate property scale (Teschner *et al.*, 1994) (Fig. 17.2.3.6). This approach, however, is still tied to the level of triangulation. The more general two-dimensional or location-based surface texture mapping requires a global scheme for assigning texture indices. While the original molecular surface geometry does not lend itself directly to this type of texture mapping, recent analytical approximations to these surfaces, such as spherical-harmonics-based molecular surfaces (Duncan & Olson, 1993), provide simple hierarchical meshing schemes that can be easily texture mapped by using a 'Mercator'-like projection between the image and the molecular surface (Duncan & Olson, 1995) (Fig. 17.2.3.6).

17.2.3.2. Volumetric representation

Molecular properties are not confined to bonds and surfaces. These, in fact, are geometric constructs or abstractions of the time-dependent volumetric characteristics of molecules. In crystallography, electron density is the primary volumetric property to be visualized. Other derived or computed volumetric properties have become important to visualize as well, especially for macromolecules and their complexes. Electrostatic potential and field gradients help establish a molecule's effect at a distance, and a variety of volumetric atomic affinity potentials or grids (Goodford, 1985) can provide a picture of the types of molecular interactions that are energetically favoured.

Traditionally, electron density and other volumetric properties have been displayed as isocontour or isosurface representations, in which lines or surfaces of constant value are rendered in planes or in 3-D space to reveal characteristics of the volumetric property. Early computer-graphic pen plots of planar Fourier projections of electron density were usually sufficient to reveal atomic structure. As the molecules of study became larger and more complex, stacks of two-dimensional slices, creating three-dimensional isocontours, became necessary. The first computer representations of such 3-D isovalue surfaces were composed of three orthogonal 2-D plots – giving the impression of a 'basket weave'. These plots depicted surface isocontours of the three-dimensional density, but had several problems from a computational and representational point of view. Since there were preferred directions of the contours (along the x, y and z axes), particular views were difficult to interpret. Additionally, the three

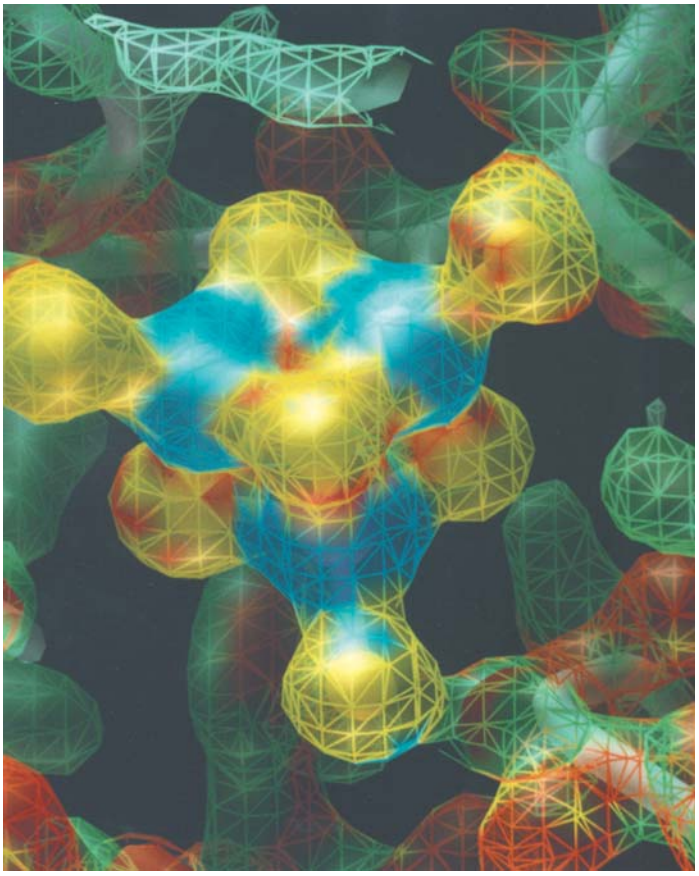


Figure 17.2.3.7
Crystallographic electron-density isosurfaces, showing details of a protein iron–sulfur cluster. The surfaces are coloured by the gradient of the electron density, highlighting the iron and sulfur densities. Image by Michael Pique, The Scripps Research Institute.

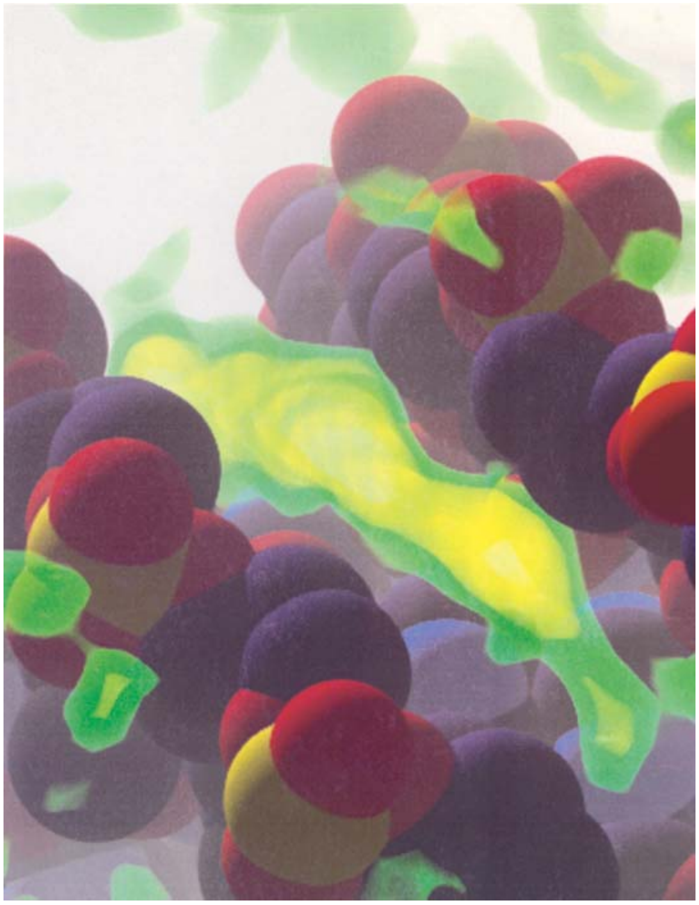


Figure 17.2.3.8
A difference-electron-density map of a minor-groove drug binding in DNA. This image combines volumetric rendering of the electron density with a geometric model of the DNA molecule. Data courtesy of R. E. Dickerson, UCLA. Image by David Goodsell, The Scripps Research Institute.

orthogonal contours did not define a well formed triangulated geometric surface, so modern surface rendering techniques could not be applied directly. Moreover, the computation and recomputation of isosurfaces was relatively inefficient. An algorithm to compute directly the three-dimensional isosurface, called 'marching cubes', was devised by Lorenson & Kline (1987) (Fig. 17.2.3.7). This algorithm speeded up the contouring process and enabled shaded surface representation of these surfaces. More recently, the re-computation of isosurfaces has been speeded up through the pre-computation of seed points that span all values of the volume. Using these seed points to flood-fill an isosurface of a given value reduces the contouring computation from a three-dimensional to a two-dimensional calculation. This enables the interactive modification of contour levels for even very large volumes (Bajaj *et al.*, 1996).

While isocontours and isosurfaces have been the dominant modes of volumetric representation in molecular graphics, there has been a trend in scientific visualization to use alternative techniques, termed 'direct volume rendering'. These methods bypass the construction of contours or surfaces to represent values within the volume, and instead use the scalar (or sometimes vector) values within the volume to produce an image directly. A general technique to accomplish this type of volumetric rendering is termed ray casting. If one considers a function that maps the scalar values of a volume into optical properties such as colour and opacity, one can simulate the passage of light rays through the volume, projecting the resulting rays onto the image plane. Given an appropriate transfer function or look-up table, the image represents the distribution of all of the values within the volume, circumventing the need to select only certain values as required for isocontouring. Such techniques have been used extensively in medical tomography (Höhne *et al.*, 1989) and electron microscopy (Kremer *et al.*, 1996; Hessler *et al.*, 1996). Their use has also been explored in the rendering of volumetric properties of molecules (Goodsell *et al.*, 1989). The images that are obtained by direct volume rendering tend to appear cloud-like, with soft edges. While this may be a 'true' representation of the molecular characteristics, it is sometimes difficult to interpret visually. Techniques for imparting shading cues into these renderings by using gradient information in the volume has made this type of rendering more interpretable (Drebein *et al.*, 1988). Another potential drawback to these methods is the cost of the computations. Since these methods require computing the effect of every element of the volume, the amount of computation scales as the cube of the linear dimension. There have been several clever software and hardware approaches to overcoming this problem. One novel hardware approach is to use three-dimensional texture mapping. By stacking texture-mapped planes to represent the colour and opacity of the volume, and using the hardware depth-buffer capabilities to compose the final image in the viewing plane, one can manipulate and render reasonable-size volumes (128^3) at highly interactive rates. For molecular visualization, one would like to be able to represent both geometric and volumetric characteristics in the same rendering to visualize, for instance, model and data (Fig. 17.2.3.8). The three-dimensional texture-mapping approach enables this easily, since the planes upon which the volume data are mapped are in fact geometric. Other direct-volume rendering codes provide this capability as well.

17.2.3.3. *Information visualization*

While molecular-structure research deals directly with objects in three dimensions, it is at times advantageous to abstract this three-dimensional information into diagrams that show relationships that are not readily apparent by examination of a set of geometric models or volumes themselves. This type of representation is broadly termed 'information visualization'. In the arena of molecular structure, probably the best known and most widely used diagram of this type is the Ramachandran plot (Ramachandran & Sasisekharan, 1968), which maps the positions of each of a protein's amino-acid residues into the backbone torsion-angle space of φ and ψ. Such a diagram readily pinpoints the parts of the protein backbone that have unusual (and sometimes erroneous) configurations. It also nicely shows the clustering of residues into the standard secondary structural motifs and their variations. There have been several enhancements of the Ramachandran plot over the years, some of which superimpose computed energy contours or colour-code residues by characteristics such as sequence order.

Another visualization approach that has become very useful is the distance matrix plot, and its derivative, the difference distance matrix (Phillips, 1970). By constructing a matrix of distances between each amino-acid α-carbon and contouring or colouring the resulting values, one can readily see the patterns of α-helices and β-sheets within the structure. An advantage of this type of visualization is that it is coordinate-frame independent. Thus two structures can be compared for features without first superposing their coordinates in the same frame. This approach also works well when comparing two different structures of the same molecule, where there may be some movement between the two. By computing the distance matrix for each structure, and then computing the difference between the two distance matrices, the resulting difference distance matrix will indicate those parts of the structure that stay in the same relative relationships and those that may move relative to each other (Fig. 17.2.3.9).

Animating trajectories of molecular structures and changes in volumetric properties over time is one way to look for trends and patterns in molecular dynamics and other time-course simulations. However, other modes of information visualization can assist analysis and communication of results, sometimes more effectively. Plotting an array of small images showing the time course of key properties can reveal patterns that may be difficult to see in a trajectory. For instance, using the program *MolMol* (Koradi *et al.*, 1996), the time course of the seven nucleic-acid backbone torsion angles during a dynamics simulation of an RNA polynucleotide can be plotted on a circular graph (starting from the centre and progressing outward) to uncover patterns of change and correlation between a large number of variables over time.

In addition to the enormous amount of information generated by computational simulations of molecular dynamics, dockings and other multi-structure, multi-modal techniques, the floodgates of molecular information have opened, gushing data from genomics and high-throughput structure determination. Thus, the need for novel visualization methods has become even more acute. Circle maps defining genomic structure at various levels of detail and annotation have become a common graphical form for organizing and communicating the positional and functional aspects of genome structure. Aligned nucleic acid or amino-acid sequences coded by conservation, chemical property, or any number of other functional relationships have become the *lingua franca* of gene hunters and gatherers. As the protein structure database continues its exponential growth, the opportunities for defining and refining structural family relationships abound. Developing methods for effectively visualizing the relationships that arise from all-by-all computational comparisons of the entire database is an important current challenge in molecular graphics.

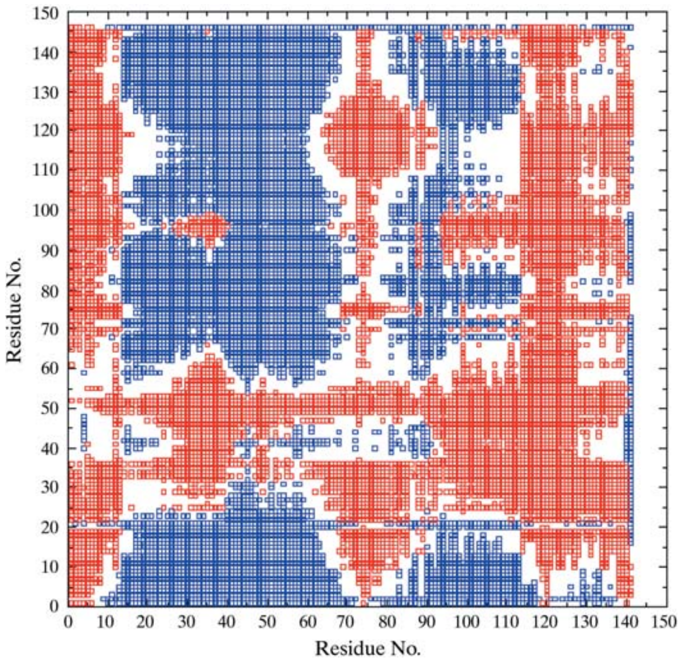
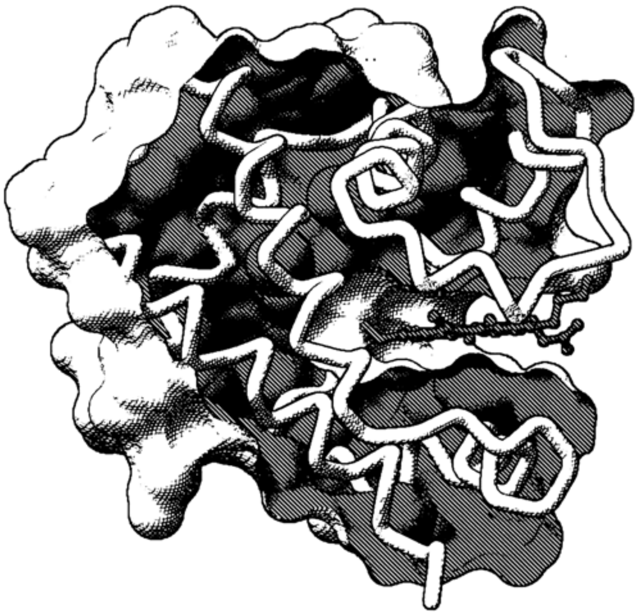


Figure 17.2.3.9
A difference distance matrix plot of the α_1–β_2 interface of haemoglobin in the T to R transformation. The *x* axis represents the α_1 subunit and the *y* axis the β_2 subunit. Red points indicate residues that are closer following the transformation and blue points indicate residues that move farther apart. Plot by Raj Srinivasan, Johns Hopkins University.

17.2.4. Presentation graphics

Much of molecular graphics can be classified as working or 'throw-away' graphics. Typically this involves the interactive creation of graphical representations on screen or paper that are used in the course of research to build, modify and analyse molecular structures, their motions and interactions. Such graphics need only be intelligible to the researchers involved. On the other hand, a presentation or publication graphic must be able to stand on its own to convey information to a broader audience. Thus it requires additional thought and work in its creation. It is unfortunate that many scientists simply capture the working graphics on their computer screens for use in publications or other forms of communication. Both interpretability and intelligibility may suffer badly if a number of issues are not considered in the production of a presentation graphic: What is the medium of publication? What is the main point of the graphic? Who are the target audience? How will reproduction or the viewing environment affect the impact of the graphic? Even seemingly simple issues such as when and how to use colour can in reality be a complex mixture of aesthetics, psychology, technology and economics. While an in-depth discussion of these issues is beyond the scope of this chapter, it is worthwhile to look at two categories of publication graphics, illustration and animation, in this context.



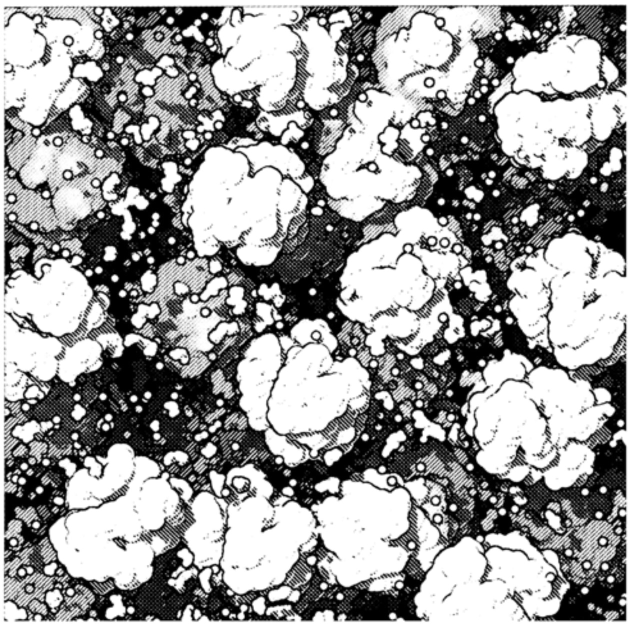


Figure 17.2.4.1
Line-art molecular illustration. The figure on the left depicts the α-carbon backbone and molecular surface of the α subunit of haemoglobin (PDB code: 2hhb). Outlines define the shapes of the surfaces and tubes, and contour lines enhance their three-dimensionality. Hatch lines, normally used for shadows, are used here to darken the inside of the molecular surface. The figure on the right shows the interior of a human red blood cell (300 Å^3), showing all molecules except water. The picture is drawn with outlines defining each molecule. Shadows and depth cuing are used to enhance the three-dimensional character of the image. Contour lines are not used. From Goodsell & Olson (1992).

17.2.4.1. Illustration

In print media, shaded colour images can present a number of difficulties. In addition to the issue of cost, colour shifting, reproducibility and loss of detail in the half-toning process may lead to less than the desired result. Simple line art is an effective way to bypass many of these complications. Since the advent of printing in the middle ages, artists and scientists have explored the problems of creating illustrations within the limitations of the printing process. Over time, artists have built a vocabulary of outlines, hatched shading and varied textures to simplify and clearly portray an object. While the creation of such illustrations was time consuming and required considerable artistic talent, they effectively portrayed the observational science of the day. The advent of computers and computer graphics removed any requirement for skilled hand draftsmanship in the production of molecular representations, but did not solve all of the problems of good illustration. As mentioned above, prior to the widespread use of interactive computer graphics, molecular structures were often published as outline drawings of ball-and-stick models using programs such as *ORTEP* or *PLUTO* (Motherwell & Clegg, 1978). More recently, programs such as *MOLSCRIPT* (Kraulis, 1991) have re-established the popularity of line-art illustration in the molecular realm. A good *ORTEP* drawing usually took a great deal of preparation time in order to get the best representation and viewpoint to display the structure effectively. As the visual repertoire of molecular structure has expanded to a wide variety of shapes including ribbons, tubes and solvent-based surfaces, the challenge of automating the general illustration process has grown. A number of techniques have been developed in the computer-graphics community to generate images in the style of technical or artistic illustrations (Fig. 17.2.4.1). These approaches use lighting, depth information and geometry to produce black-and-white drawings with shapes defined by silhouette lines and cross-hatched shading, and details shown by a variety of textures. *MOLSCRIPT* has used some of these techniques for ribbon and ball-and-stick renderings. More general applications of these approaches to molecular illustration have also been described by Goodsell & Olson (1992).

A significant advantage of digital black-and-white illustration is the efficiency of representation. Since each picture element takes only a single bit of information (black or white), and since there are typically large areas that are of constant value, these

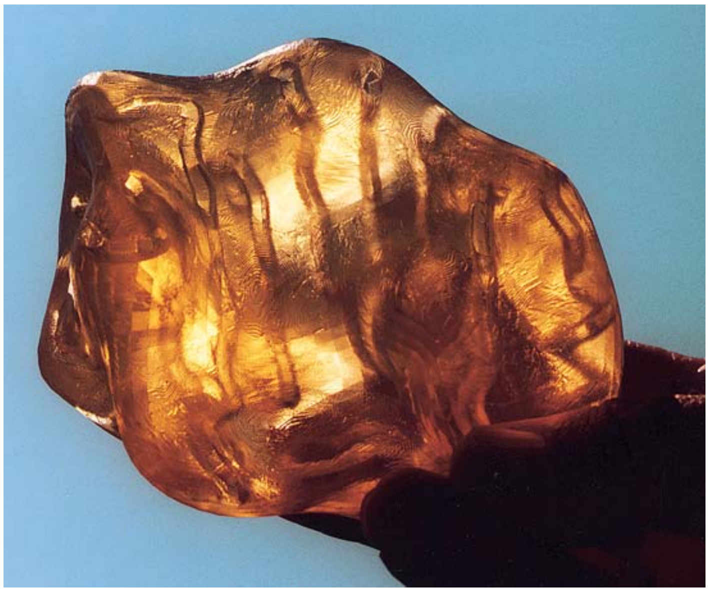


Figure 17.2.4.2
A stereolithographic model of betalactamase inhibitory protein (BLIP). The model depicts the shape of the molecule as represented by a spherical-harmonic approximation of the solvent-excluded surface. Inside the surface is a hollow tube which follows the α-carbon trace of the protein backbone. The model was designed by Michael Pique and Arthur Olson, and fabricated on a 3D systems SLA 1000 by Beryl Hodgson.

images can be compressed, stored, transmitted and printed very efficiently. Thus, with the advent of electronic web publication, such illustrations represent an attractive alternative to full colour. These same characteristics represent significant advantages for the digital transmission and use of animated sequences as well.

17.2.4.2. Animation

Computer-graphic molecular animations began to appear in the late 1960s. Recording directly off their vectorscope, Levinthal and colleagues in Project Mac produced a record of an interactive molecular modelling session in 1967. In the early 1970s, a number of molecular animations were produced to convey new scientific results. Wilson at UC San Diego showed vibrational modes of small molecules in a film produced frame-by-frame on a vectorscope. Parr & Polyani painstakingly filmed pen plotter drawings of space-filling diatomic molecules to animate a bimolecular chemical reaction. Sussman & Seeman produced a black-and-white vector animation of the dinucleotide UpA structure in 1972 by recording directly off a vectorscope. Seeman, Rosenberg & Meyserth produced a more ambitious molecular animation in 1973 entitled *Deep Groove*, which depicted the structure of double helical segment ApU CpC and its implications for more extended DNA geometry. This film was shot in colour, using a monochrome vectorscope and multiple exposures through a colour filter wheel. Around this time, Knowlton, Cherry & Gilmer at Bell Labs used early frame-buffer devices to display and animate patterns of crystal growth based upon aggregation of spheres. In the mid-1970s, Porter & Feldman had developed a scan-line based CPK representation for raster displays and had animated molecular structures, and Langridge and co-workers had taken up recording off the black-and-white vector displays then available. By the end of the 1970s, Max had produced high-quality animations of DNA using a high-resolution Dicomed film writer (Max, 1983) and Olson had used an early colour vector display from Evans and Sutherland to produce an eight-minute animation depicting the structure of tomato bushy stunt virus (Olson, 1981). By the early 1980s, animation projects became more ambitious. Olson produced large-screen OmniMax DNA and virus animation segments for Disney's EPCOT center in 1983. Max produced a red–blue stereo OmniMax film for Fujitsu entitled *We Are Born of Stars*, which included a continuous scene depicting the hierarchical packaging of DNA from atoms to chromosomes, based on the best current model of the time.

Computer-graphics animation has presented both great potential and significant challenges to the molecular scientist wishing to communicate the results of structural research. Animation can not only enhance the depiction of three-dimensional structure through motion stereopsis, it can show relationships through time, and demonstrate mechanism and change. The use of pans, zooms, cuts and other film techniques can effectively lead the viewer through a complex scene and focus attention on specific structures or processes. The vocabulary of film, video and animation is familiar to all, but can be a difficult language to master. While short animations showing simple rotations or transitions between molecular states, or dynamics trajectories, are now routinely made for video or web viewing, extended animations showing molecular structure and function in depth are still relatively rare. The time, tools and expertise that are required are not generally available to structural researchers.

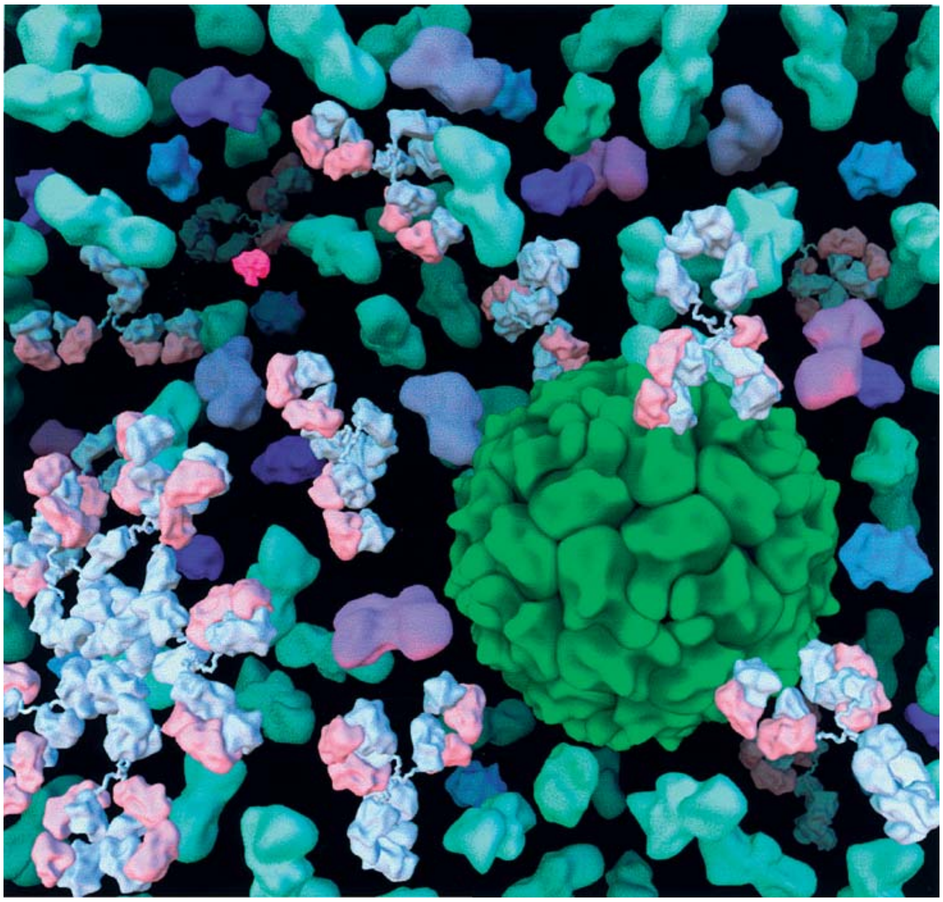


Figure 17.2.5.1

This image represents a volume of blood plasma 750 Å on a side. Within the three-dimensional model, antibodies (Y- and T-shaped molecules in light blue and pink) are binding to a virus (the large green spherical assembly on the right), labelling it for destruction. It shows all macromolecules present in the blood plasma at a magnification of about 10 000 000 times. This model is composed of over 450 individual protein domains, ranging in size from the 60 protomers making up the poliovirus to a single tiny insulin molecule (in magenta). The model was constructed using atomic level descriptions for each molecule, for a total of roughly 1.5 million atoms. Detailed surfaces were computed for each type of protein using *MSMS* by Michel Sanner and then smoothed to a lower resolution using the *HARMONY* spherical-harmonic surfaces developed by Bruce Duncan. The model geometry contains over 1.5 million triangles.

17.2.4.3. The return of physical models

While the use of physical models of molecules has largely been replaced by computer graphics, new computer-driven rapid-prototyping technologies which originated in the manufacturing sector have begun to be utilized in the display of molecular structure. A number of 'three-dimensional printing' methods have been developed to build up a physical model directly from a computational surface representation of an object (Burns, 1954). One of the earliest methods, stereolithography, uses a resin which is polymerized when exposed to laser light of a given wavelength. The laser is passed through a vat of the liquid resin and is lowered, layer-by-layer as it plots out the shape of the object (Fig. 17.2.4.2). Other approaches build up layers of paper or plastic through lamination or deposition. These methods have been used by a number of scientists to produce various representations of molecular structure (Bailey *et al.*, 1998). The ability to hold an accurate representation of a molecular surface in one's hand and feel its shape can give great insight, not only to people with visual impairments, but to anyone. Moreover, when one is dealing with processes such as docking and assembly, these physical models can add a haptic and manipulative appreciation of the nature of the problem. While at this point colour has not been implemented in these technologies, there remains the promise that such automated production of molecular models will enhance the communication and appreciation of molecular structure.

17.2.5. Looking ahead

Moore's law has already delivered on the promise of three-dimensional graphics capability for the desktop and laptop. The internet and World Wide Web have made molecular structure data and display software available to the masses. Have molecular graphics reached a stage of maturity beyond which only small incremental changes will be made?

The Human Genome Initiative and high-throughput structure determination are beginning to change the scope of the questions asked of molecular modelling. Prediction of function, interactions, and large-scale assembly and mechanism will become the dominant domain of molecular graphics and modelling. These tasks will challenge the capabilities of the hardware, software and, particularly, the user interface. New modes of interacting with data and models are coming from the computer-graphics community. Molecular docking and protein manipulation using force-feedback devices have been demonstrated at the University of North Carolina (Brooks *et al.*, 1990). The same team has developed a 'nanomanipulator' which couples a scanning atomic force microscope with stereoscopic display and force-feedback manipulation to control and sense the positioning and interactions of the probe within the molecular landscape (Taylor *et al.*, 1993). The challenge of bridging across the scales of size and complexity of the molecular world may lead us into the realm of virtual reality. Data from X-ray crystallography are being combined with data from large molecular complexes, characterized by electron microscopy. These data, in turn, can be integrated with those from optical confocal microscopy and other imaging techniques. With structures of molecules, assemblies and distributions, as well as data on molecular inventories, we can start to piece together integrated pictures of cellular environments, but with full atomic modelling at the base (Fig. 17.2.5.1). Thus, while climbing around inside a protein molecule might not add much in the way of perceptual advantage, navigating through the molecular environment of a cell may prove to be instructive as well as inspirational.

References

Bailey, M., Schulten, K. & Johnson, J. (1998). *The use of solid physical models for the study of macromolecular assembly. Curr. Opin. Struct. Biol.* **8**, 202–208.

Bajaj, C. L., Pascucci, V. & Schikore, D. R. (1996). *Fast isocontouring for improved interactivity*. In *Proceedings of the ACM SIGGRAPH/IEEE Symposium on Volume Visualization*, pp. 39–46. San Francisco: ACM Press.

Barry, C. D. & McAlister, J. P. (1982). *High performance molecular graphics, a hardware review*. In *Computational Crystallography*, edited by D. Sayre, pp. 274–255. Oxford: Clarendon Press.

Branden, C.-I., Jornvall, H., Eklund, H. & Fureugren, B. (1975). *Alcohol dehydrogenase*. In *The Enzymes*, edited by P. Boyer, pp. 104–186. New York: Academic Press.

Brooks, F. P. Jr, Ouh-Young, M., Batter, J. J. & Kilpatrick, P. J. (1990). *Project GROPE – haptic displays for scientific visualization. Comput. Graphics*, **24**, 177–186.

Burns, M. (1954). *Automated Fabrication: Improving Productivity in Manufacturing*. New Jersey: Prentice Hall.

Carson, M. (1991). *RIBBONS 2.0. J. Appl. Cryst.* **24**, 958–961.

Connolly, M. L. (1983). *Solvent-accessible surfaces of proteins and nucleic acids. Science*, **221**, 709–713.

Connolly, M. L. & Olson, A. J. (1985). *GRANNY, a companion to GRAMPS for the real-time manipulation of macromolecular models. Comput. Chem.* **9**, 1–6.

Diamond, R. (1982). *BILDER: an interactive graphics program for biopolymers*. In *Computational Crystallography*, edited by D. Sayre, pp. 318–325. Oxford: Clarendon Press.

Drebein, R., Carpenter, L. & Hanrahan, P. (1988). *Volume rendering*. Proc. ACM SIGGRAPH'88 (Atlanta, Georgia, 1–5 August 1988). In *Comput. Graphics Proc. Annu. Conf. Ser. 1988* (1993), pp. 65–74. New York: ACM SIGGRAPH.

Duncan, B. S. & Olson, A. J. (1993). *Approximation and characterization of molecular surfaces. Biopolymers*, **33**, 219–229.

Duncan, B. S. & Olson, A. J. (1995). *Approximation and visualization of large-scale motion of protein surfaces. J. Mol. Graphics*, **13**, 250–257.

Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988). *The MIDAS display system. J. Mol. Graphics*, **6**, 13–27.

Goodford, P. J. (1985). *A computational procedure for determining energetically favorable binding sites on biologically important molecules. J. Med. Chem.* **28**, 849–857.

Goodsell, D. S., Mian, I. S. & Olson, A. J. (1989). *Rendering of volumetric data in molecular systems. J. Mol. Graphics*, **7**, 41–47.

Goodsell, D. S. & Olson, A. J. (1992). *Molecular illustration in black and white. J. Mol. Graphics*, **10**, 235–240.

Gouraud, H. (1971). *Continuous shading of curved surfaces. IEEE Trans. Comput.* **20**, 623–628.

Gwilliam, M. & Max, N. (1989). *Atoms with shadows – an area-based algorithm for cast shadows on space-filling molecular models. J. Mol. Graphics*, **7**, 54–59.

Hessler, D. S., Young, S. J. & Ellisman, M. H. (1996). *A flexible environment for visualization of three-dimensional biological structures. J. Struct. Biol.* **116**, 113–119.

Höhne, K. H., Bomans, M., Pommert, A., Reimer, M., Schiers, C., Tiede, U. & Wiebecke, G. (1989). *3D-Visualization of Tomographic Volume Data Using the Generalized Voxel-Model*. Volume visualization workshop, Chapel Hill, NC. Department of Computer Science, University of North Carolina at Chapel Hill.

Hubbard, R. E. (1986). *HYDRA: current and future developments*. In *Computer Graphics and Molecular Modelling*, edited by R. Fletterick & M. Zoller, pp. 9–12. Cold Spring Harbor Press.

Johnson, C. K. (1970). *ORTEP: a Fortran Thermal-Ellipsoid Plot Program for Crystal Structure Illustrations*. Report ORNL 3794. Oak Ridge National Laboratory, Tennessee, USA.

Jones, T. A. (1978). *A graphics model building and refinement system for macromolecules. J. Appl. Cryst.* **11**, 268–272.

Kabsch, W., Mannherz, H. G., Suck, D., Pai, E. F. & Holmes, K. C. (1990). *Atomic structure of the actin/DNAse I complex. Nature (London)*, **347**, 37–44.

Koradi, R., Billeter, M. & Wuthrich, K. (1996). *MOLMOL: a program for display and analysis of macromolecular structures. J. Mol. Graphics*, **14**, 51–55.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst.* **24**, 946–950.

Kremer, J. R., Mastronarde, D. N. & McIntosh, J. R. (1996) *Computer visualization of three-dimensional image data using IMOD. J. Struct. Biol.* **116**, 71–76.

Lauher, J. W. (1990). *Chem-Ray: a molecular graphics program featuring an umbra and penumbra shadowing routine. J. Mol. Graphics*, **8**, 34–38.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol.* **55**, 379–400.

Levinthal, C. (1966). *Molecular modeling by computer. Sci. Am.* **214**, 42–52.

Lorenson, W. E. & Kline, H. E. (1987). *Marching cubes: a high resolution 3D surface construction algorithm. Comput. Graphics*, **21**, 163–169.

Max, N. (1983). *SIGGRAPH'84 call for OmniMax films. Comput. Graphics*, **17**, 73–76.

Motherwell, W. D. S. & Clegg, W. (1978). *PLUTO. Program for Plotting Molecular and Crystal Structures*. University of Cambridge, England.

Mujeeb, A., Kerwin, S. M., Kenyon, G. L. & James, T. L. (1993). *Solution structure of a conserved DNA sequence from the HIV-1 genome: restrained molecular dynamics simulation with distance and torsion angle restraints derived from two-dimensional NMR spectra. Biochemistry*, **32**, 13419–13431.

Nichols, W. L., Rose, G. D., Ten Eyck, L. F. & Zimm, B. H. (1995). *Rigid domains in proteins: an algorithmic approach to their identification. Proteins Struct. Funct. Genet.* **23**, 38–45.

O'Donnell, T. J. & Olson, A. J. (1981). *GRAMPS – a graphics language interpreter for real-time, interactive three-dimensional picture editing and animation. Comput. Graphics*, **15**, 133–142.

Olson, A. J. (1981). *Tomato Bushy Stunt Virus.* Film. Lawrence Berkeley Laboratory, University of California, Berkeley, California, USA.

Olson, A. J. (1983). *Computer graphics in biomolecular science*. NICOGRAPH '83, pp. 332–356. Tokyo: Nihon Keizai Shimbun, Inc.

Pepinsky, R. (1952). *X-RAC and S-FAC: electronic analogue computers for X-ray analysis.* In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*. Pennsylvania State College, USA.

Phillips, D. C. (1970). *British Biochemistry, Past and Present*, edited by T. W. Goodwin, pp. 11–28. Academic Press.

Phong, B. T. (1975). *Illumination for computer generated images. Commun. ACM*, **18**, 311–317.

Porter, T. K. (1979). *The shaded surface display of large molecules. Comput. Graphics*, **13**, 234–236.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Conformation of polypeptides and proteins. Adv. Protein Chem.* **23**, 283–437.

Richards, F. M. (1977). *Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. A*, **6**, 151–176.

Richardson, J. S. (1981). *The anatomy and taxonomy of protein structure. Adv. Protein Chem.* **34**, 167–339.

Sanner, M.-F., Olson, A. J. & Spehner, J.-C. (1996). *Reduced surface: an efficient way to compute molecular surfaces. Biopolymers*, **38**, 305–320.

Sayle, R. A. & Milner-White, E. J. (1995). *RASMOL: biomolecular graphics for all. Trends Biochem. Sci.* **20**, 374.

Taylor, R. M. I., Robinett, W., Chi, V. L., Brooks, F. P. Jr, Wright, W. V., Williams, R. S. & Snyder, E. J. (1993). *The nanomanipulator: a virtual reality interface for a scanning tunnelling microscope*. Proc. ACM SIGGRAPH'93 (Anaheim, California, 1–6 August 1993). In *Comput. Graphics Proc. Annu. Conf. Ser. 1993* (1993), pp. 127–133. New York: ACM SIGGRAPH.

Teschner, M., Henn, C., Vollhardt, H., Reiling, S. & Brinkmann, J. (1994). *Texture mapping: a new tool for molecular graphics. J. Mol. Graphics*, **12**, 98–105.

Wright, W. V. (1982). *GRIP – an interactive computer graphics system for molecular studies*. In *Computational Crystallography*, edited by D. Sayre, pp. 294–302. Oxford: Clarendon Press.

references

International Tables for Crystallography (2012). Vol. F, Chapter 18.1, pp. 459–465.

PART 18. REFINEMENT

Chapter 18.1. Introduction to refinement

L. F. Ten Eyck and K. D. Watenpaugh

18.1.1. Overview

Methods of improving and assessing the accuracy of the positions of atoms in crystals rely on the agreement between the observed and calculated diffraction data. Calculation of diffraction data from an atomic model depends on the theoretical model of scattering of X-rays by crystals discussed in *IT* B (2008) Chapter 1.2. The properties of the measured data are discussed in *IT* C (2004) Chapters 2.2 and 7.1–7.5, and the mathematical basis of refinement of structural parameters is discussed in *IT* C Chapters 8.1–8.5. This chapter concentrates on the special features of macromolecular crystallography.

18.1.2. Background

Macromolecular crystallography is not fundamentally different from small-molecule crystallography, but is complicated by the sheer size of the problems. Typical macromolecules contain thousands of atoms and crystallize in unit cells of around a million cubic ångstroms. The large size of the problems has meant that the techniques applied to small molecules require too many computational resources to be directly applied to macromolecules. This has produced a lag between macromolecular and small-molecule practice beyond the limitations introduced by generally poorer resolution. In essence, macromolecular refinement has followed small-molecule crystallography. Additional complexity arises from the book-keeping required to describe the macromolecular structure, which is usually beyond the capabilities of programs designed for small molecules.

Fitting the atom positions to the calculated electron-density maps (Fourier maps) was a standard method until the introduction of least-squares refinement technique in reciprocal space by Hughes (1941). A less computationally intense method of calculating shifts using difference Fourier maps (ΔF methods) was introduced by Booth (1946*a*,*b*). By the early 1960s, digital computers were becoming generally available and least-squares refinement methods became the method of choice in refining small molecules. The program *ORFLS* developed by Busing *et al.* (1962) was perhaps the most extensively used. In the late 1960s, as protein structures were being determined by multiple isomorphous replacement (MIR) methods (see Part 12 and Chapter 2.4 in *IT* B), methods of improving the structural models derived from the electron-density maps were being studied. Diamond (1971) introduced the use of a constrained chemical model in the fitting of a calculated electron-density model to an MIR-derived electron-density map in a 'real-space refinement' procedure. Diamond commented that phases derived from a previous cycle of real-space fitting could be used to calculate the next electron-density map, but this was not done. Watenpaugh *et al.* (1972) first showed in 1971 that ΔF refinement methods could be applied to both improve the model and extend the phases from initial MIR or SIR (single isomorphous replacement) experimental phases. Watenpaugh *et al.* (1973) also applied least-squares techniques to the refinement of a protein structure for the first time using a 1.54 Å resolution data set. Improvement of the phases, clarification of the electron-density maps and interpretation of unknown sequences in the structure were clearly evident, although chemical restraints were not applied. The adaptation by Hendrickson & Konnert of the restrained least-squares refinement program developed by Konnert (Konnert, 1976; Hendrickson, 1985) became the first extensively used macromolecular refinement program. At this time, refinement of protein models became practical and nearly universal. Model refinement improved models derived from structures determined by isomorphous replacement methods and also provided the means to improve structural models of related protein structures determined by molecular replacement methods (see Part 13 and *IT* B Chapter 2.3).

By the 1980s, it became clear that additional statistical rigour in macromolecular refinement was required. The first and most obvious problem was that macromolecular structures were often solved with fewer observations than there were parameters in the model, which leads to overfitting. Recent advances include cross validation for detection of overfitting of data (Brünger, 1992); maximum-likelihood refinement for improved robustness (Pannu & Read, 1996; Murshudov *et al.*, 1997; Bricogne, 1997; Adams *et al.*, 1999); improved methods for describing the model with fewer parameters (Rice & Brünger, 1994; Murshudov *et al.*, 1999); and incorporation of phase information from multiple sources (Pannu *et al.*, 1998). These improvements in the theory and practice of macromolecular refinement will undoubtedly not be the last word on the subject.

18.1.3. Objectives

A variety of methods are employed to improve the agreement between observed and calculated macromolecular diffraction patterns. Some of the more popular methods are discussed in the different sections of this chapter. In part, the different methods arise from focusing on different goals during different stages of model refinement. Bias generated by incomplete models, and radius of convergence, are important considerations at early stages of refinement, because the models are usually incomplete, contain significant errors in atom parameters and may carry errors from misinterpretation of poorly phased electron-density maps. During this stage of the process, the primary concern is to determine how the model of the chain tracing and conformation of the residues should be described. In later stages, after the description of the model has been determined, the objective is to determine accurate estimates of the values of the parameters which best explain the observed data. These two stages of the problem have different properties and should be treated differently.

Copyright © 2012 International Union of Crystallography

18.1.4. Least squares and maximum likelihood

'Improving the agreement' between the observed and calculated data can only be done if one first decides the criteria to be used to measure the agreement. The most commonly used measure is the L_2 norm of the residuals, which is simply the sum of the squares of the differences between the observed and calculated data (*IT* C Chapter 8.1),

$$L_2(\mathbf{x}) = \|w_i[y_i - f_i(\mathbf{x})]\| = \sum_i w_i[y_i - f_i(\mathbf{x})]^2, \qquad (18.1.4.1)$$

where w_i is the weight of observation y_i and $f_i(\mathbf{x})$ is the calculated value of observation i given the parameters $\mathbf{x}$. In essence, least-squares refinement poses the problem as 'Given these data, what are the parameters of the model that give the minimum variance of the observations?'. The L_2 norm is strongly affected by the largest deviations, which is not a desirable property in the early stages of refinement where the model may be seriously incomplete. In the early stages, it may be better to refine against the L_1 norm,

$$L_1 = \sum_i w_i|[y_i - f_i(\mathbf{x})]|,$$

the sum of the absolute value of the residuals. At present, this technique is not used in macromolecular crystallography.

The observable quantity in crystallography is the diffracted intensity of radiation. Fourier inversion of the model gives us a complex structure factor. The phase information is normally lost in the formulation of $f_i(\mathbf{x})$. This is a root cause of some of the problems of least-squares refinement from poor starting models. Many of the problems of least-squares refinement can be addressed by changing the measure of agreement from least squares to maximum likelihood, which evaluates to the likelihood of the observations given the model. In this formulation, the problem is posed as 'Given this model, what is the probability that the given set of data would be observed?'. The model is adjusted to maximize the probability of the given observations. This procedure is subtly different from least squares in that it is reasonably straightforward to account for incomplete models and errors in the model in computing the probability of the observations. Maximum-likelihood refinement is particularly useful for incomplete models because it produces residuals that are less biased by the current model than those produced by least squares. Maximum likelihood also provides a rigorous formulation for all forms of error in both the model and the observations, and allows incorporation of additional forms of prior knowledge (like additional phase information) into the probability distributions.

The likelihood of a model given a set of observations is the product of the probabilities of all of the observations given the model. If $P_a(\mathbf{F}_i;\ \mathbf{F}_{i,c})$ is the conditional probability distribution of the structure factor $\mathbf{F}_i$ given the model structure factor $\mathbf{F}_{i,c}$, then the likelihood of the model is

$$L = \prod_i P_a(\mathbf{F}_i;\ \mathbf{F}_{i,c}).$$

This is usually transformed into a more tractable form by taking the logarithm,

$$\log L = \sum_i \log P_a(\mathbf{F}_i;\ \mathbf{F}_{i,c}).$$

Since the logarithm increases monotonically with its argument, the two versions of the equation have maxima at the same values of the parameters of the model. This formulation is described in more detail in Chapter 18.2, in *IT* C Section 8.2.1 and by Bricogne (1997), Pannu & Read (1996), and Murshudov *et al.* (1997).

18.1.5. Optimization

Once the choice of criteria for agreement has been made, the next step is to adjust the parameters of the model to minimize the disagreement (or maximize the agreement) between the model and the data. The literature on optimization in numerical analysis and operations research, discussed in *IT* C Chapters 8.1–8.5, is very rich. The methods can be characterized by their use of gradient information (no gradients, first derivatives, or second derivatives), by their search strategy (none, downhill, random, annealed, or a combination of these), and by various performance measures on different classes of problems. These will be discussed more fully in Section 18.1.8.

18.1.6. Data

Resolution, accuracy, completeness and weighting of data all have an impact on the refinement process. Small-molecule crystals usually, but not always, diffract to well beyond atomic resolution. Macromolecular crystals do not generally diffract to atomic resolution. Macromolecular structures are by definition large, which in turn means that the unit cells are large and the number of diffracting unit cells per crystal is small when compared to small-molecule crystals of similar size. Fortunately, the situation can be partially offset with the use of the much more intense radiation generated by synchrotrons (Part 8) and by improved data-collection methods (Parts 7–11). Synchrotron-radiation sources designed to produce intense beams of X-rays for the study of materials are becoming much more readily available. As a consequence, both higher resolution and statistically better data can be obtained. Improvements in area-detector technology, protein purification, cryocrystallography and data-integration software beneficially influence the refinement process.

Refinement of crystal structures is a statistical process. There is no substitute for adequate amounts of accurate, correctly weighted data. Lower accuracy can be accommodated by increased amounts of data and correct weighting. Unfortunately, determining the correct weighting for macromolecular diffraction data is difficult. Maximum-likelihood methods are more robust than least-squares methods against improperly weighted data.

It has been clearly demonstrated that the best procedure for refining small molecules is to include all of the observations as integrated intensities, properly weighted, without preliminary symmetry averaging. Inclusion of weak data and refinement on diffracted intensity does not change the results very much, but has a strong effect on the precision of the parameter estimates derived from the refinement.

The long-standing debate as to whether refinement should be against structure-factor amplitudes or diffracted intensity has been resolved for small-molecule crystallography. Refinement against intensity is preferred because it is closer to the experimentally observed quantity, and the statistical weighting of the data is superior to that obtained for structure-factor amplitudes. If the model is correct and the data are reasonably good, the primary distinction between the two approaches is in the standard uncertainties of the derived parameters, which are usually somewhat better if the refinement is against diffracted intensity.

18.1.7. Models

Atomic resolution models are generally straightforward. A reasonably well phased diffraction pattern at atomic resolution shows the location of each atom. The primary problem (which can be substantial) is deciding how to model any disorder that may be present. Structural chemistry is derived from the model. Macromolecular models generally have most of the structural chemistry built in as part of the model. This approach is required as a direct consequence of having too little data at too limited a resolution to determine the positions of all of the atoms without using this additional information.

There are two procedures for building structural chemistry into a model. The first is to use known molecular geometry to reduce the number of variables. For example, if the distance between two atoms is held constant, the locus of possible positions for the second atom is the surface of a sphere centred on the first atom. This means that the position of the second atom can be specified given the position of the first atom and two variables to locate the point on the sphere – a total of five variables instead of six. Every non-redundant constraint reduces the number of degrees of freedom in the model by one. If the second atom in this example were replaced by a group of atoms with known geometry (*e.g.* a phenyl group containing six atoms), the number of positional parameters could be reduced from 21 to eight. Constrained refinement is discussed extensively in *IT* C Chapter 8.3.

The second procedure is to treat the additional information as additional observations. A bond length is assumed to be an observation, based on other crystal structures, which has a mean value and a variance. This observation is added to the data instead of being used to reduce the number of parameters in the model.

The two approaches have different consequences on the ratio of observations to parameters. If we have N_o observations, N_p parameters and N_r non-redundant geometric features to add to the problem, we have either $C = N_o/(N_p - N_r)$ and $(\mathrm{d}C/\mathrm{d}N_r) = C/(N_p - N_r)$ or $C = (N_o + N_r)/N_p$ and $(\mathrm{d}C/\mathrm{d}N_r) = 1/N_p$, where C is the ratio of observations to parameters. The former are parameter *constraints* and the latter are parameter *restraints*. Constraints are more effective at increasing the ratio of observations to parameters, but since these features are built into the model, it is difficult to evaluate how appropriate they actually are for the problem at hand. Restraints provide an automatic evaluation of the appropriateness of the assumed geometry to the current data, because the deviations from the assumed values can be tested for statistical significance.

The most common constraints and restraints applied to macromolecular crystal structures are those which preserve or reinforce the molecular geometry of the amino acid or nucleotide residues (Chapter 18.3). Expected values for the geometry of these structural fragments are available from the small-molecule crystallographic literature and databases. A further step, which reduces the parameter count substantially, is to treat parts of the molecule as a set of linked rigid groups. This is particularly appropriate for aromatic fragments such as the side chains of phenylalanine, tyrosine, tryptophan and histidine, but can also be appropriate for small groups like valine and threonine. The extreme form of this approach is torsion-angle dynamics (Rice & Brünger, 1994), in which the only variables are torsion angles about bonds, and the position and orientation of the whole molecule. This description of the model works well with the right kind of optimization procedure.

Positional restraints can be parameterized in a variety of ways. For example, the geometry of three atoms can be treated as the three distances involved or as two distances and the angle between them. Several of the more popular restrained refinement programs treat the parameters for bond distances, bond angles and planarity as distances with a set of standard deviations. Others treat them as bond distances, bond angles and torsion angles weighted by the energy terms derived from experimental conditions. Different methods of parameterization and weighing have different effects on the refinement process, but to date these differences are not well characterized. The primary effects should be on the approach to convergence, as all of these formulations are normally satisfied by correct structures.

Additional criteria can be added to the model besides simple geometry. Preservation of bond lengths is usually done by adding terms

$$\sum_{\substack{\text{bonded}\\\text{atoms}}} \left(1/\sigma_{ij}^2\right)\left(d_{ij} - d_{ij}^o\right)^2$$

to the objective function, where d_{ij} is the distance between atoms i and j, d_{ij}^o is the ideal bond length, and σ_{ij} is the weight applied to the bond. This is formally equivalent to treating bond stretching as a spring. Additional energy parameters can be added, such as electrostatic energy terms. Whatever vision of reality is applied to the objective function becomes part of the model.

The atomic displacement factors (B factors) present a different set of problems from the coordinates. The behaviour of these parameters is strongly affected by coordinate errors, and in fact large atomic displacement parameters are frequently used to determine which parts of a structure are likely to contain errors. The B factors are strongly related to the rate at which the diffraction pattern diminishes with resolution and thus cannot be accurately determined unless the diffraction pattern has been measured over a sufficiently wide range of resolution to determine this rate. As a practical matter, it is not feasible to refine individual atomic displacement parameters at resolution less than about 2 Å, and they frequently present problems even in atomic resolution small-molecule structures. In high-resolution small-molecule structures, B factors are frequently represented as anisotropic ellipsoids described by six parameters per atom. In spite of the larger displacements found in macromolecules relative to small molecules, it is rarely possible to support the number of parameters required to refine a structure with independent anisotropic displacement factors. Nevertheless, the B factors of the atoms are essential parts of the crystallographic model. Several methods for reducing the number of independent B factors have been developed. The simplest is group B factors, in which one parameter is refined for all atoms in a particular group of atoms. Another method is to apply a simple model to the change in displacement parameter within a group of atoms. In this treatment, a B factor is refined for one atom, say the C_α atom of an amino-acid residue, and the remainder of the atoms in the residue are assigned displacement parameters that depend on their distance from the C_α atom (Konnert & Hendrickson, 1980). A third method is to enforce similarity of displacement parameters based on the correlation coefficients between pairs of displacement parameters in highly refined high-resolution structures (Tronrud, 1996).

Small-molecule refinement programs also apply restraints to the displacement parameters. The SIMU command of *SHELX* restrains the axes of the anisotropic displacement parameters of bonded atoms to be similar. This approach has been applied to a number of very high resolution macromolecular refinements.

Large B factors do not represent large thermal motions of the atom but rather a distribution of positions occupied by the atom over time or in different unit cells of the crystal. The line between describing atoms with large B factors as distributed about a single point or several points (disordered atoms) is sometimes blurred. At some point, the disorder can become resolved into alternative positions or the atoms disappear from the observable electron density. There are two kinds of disorder that can be easily modelled if data are available to sufficient resolution:

(1) *Static disorder* describes the situation in which portions of the structure have a small number of possible alternative conformations. The atoms in any given unit cell are in only one of the possible conformations, but different cells may have different conformations. Since the diffraction experiment averages the structure over all unit cells in the X-ray beam, the observations correspond to an average structure in which each conformation is weighted according to the fraction of the unit cells containing that conformation. The normal bond-length and angle restraints apply to each conformation, and the fractional occupancy of all conformations should sum to 1.0.

(2) *Dynamic disorder* describes the situation in which portions of the structure are not in fixed positions. This form of disorder is frequently encountered in amino-acid side chains on the molecular surface. The electrons are spread over a sufficiently large volume that the average electron density is very low and the atoms are essentially invisible to X-rays. In such cases, the best model is to simply omit the atoms from the diffraction calculation. They are commonly placed in the model in plausible positions according to molecular geometry, but this can be misleading to people using the coordinate set. If the atoms are included in the model, the atomic displacement parameters generally become very large, and this may be an acceptable flag for dynamic disorder. The hazard with this procedure is that including these atoms in the model provides additional parameters to conceal any error signal in the data that might relate to problems elsewhere in the model.

At high resolution, it is sometimes possible to model the correlated motion of atoms in rigid groups by a single tensor that describes translation, libration and screw. This is rarely done for macromolecules at present, but may be an extremely accurate way to model the behaviour of the molecules. The recent development of efficient anisotropic refinement methods for macromolecules by Murshudov *et al.* (1999) will undoubtedly produce a great deal more information about the modelling of dynamic disorder and anisotropy in macromolecular structures.

Macromolecular crystals contain between 30 and 70% solvent, mostly amorphous. The diffraction is not accurately modelled unless this solvent is included (Tronrud, 1997). The bulk solvent is generally modelled as a continuum of electron density with a high atomic displacement parameter. The high displacement parameter blurs the edges, so that the contribution of the bulk solvent to the scattering is primarily at low resolution. Nevertheless, it is important to include this in the model for two reasons. First, unless the bulk solvent is modelled, the low-resolution structure factors cannot be used in the refinement. This has the unfortunate effect of rendering the refinement of *all* of the atomic displacement parameters ill-determined. Second, omission or inaccurate phasing of the low-resolution reflections tends to produce long-wavelength variations in the electron-density maps, rendering them more difficult to interpret. In some regions, the maps can become overconnected, and in others they can become fragmented.

18.1.8. Optimization methods

Optimization methods for small molecules are straightforward, but macromolecules present special problems due to their sheer size. The large number of parameters vastly increases the volume of the parameter space that must be searched for feasible solutions and also increases the storage requirements for the optimization process. The combination of a large number of parameters and a large number of observations means that the computations at each cycle of the optimization process are expensive.

Optimization methods can be roughly classified according to the order of derivative information used in the algorithm. Methods that use no derivatives find an optimum through a search strategy; examples are Monte Carlo methods and some forms of simulated annealing. First-order methods compute gradients, and hence can always move in a direction that should reduce the objective function. Second-order methods compute curvature, which allows them to predict not only which direction will reduce the objective function, but how that direction will change as the optimization proceeds. The zero-order methods are generally very slow in high-dimensional spaces because the volume that must be searched becomes huge. First-order methods can be fast and compact, but cannot determine whether or not the solution is a true minimum. Second-order methods can detect null subspaces and singularities in the solution, but the computational cost grows as the cube of the number of parameters (or worse), and the storage requirements grow as the square of the number of parameters – undesirable properties where the number of parameters is of the order of 10^4.

Historically, the most successful optimization methods for macromolecular structures have been first-order methods. This is beginning to change as multi-gigabyte memories are becoming more common on computers and processor speeds are in the gigahertz range. At this time, there are no widely used refinement programs that run effectively on multiprocessor systems, although there are no theoretical barriers to writing such a program.

18.1.8.1. Solving the refinement equations

Methods for solving the refinement equations are described in *IT* C Chapters 8.1 to 8.5 and in many texts. Prince (1994) provides an excellent starting point. There are two commonly used approaches to finding the set of parameters that minimizes equation (18.1.4.1). The first is to treat each observation separately and rewrite each term of (18.1.4.1) as

$$w_i[y_i - f_i(\mathbf{x})] = w_i \sum_{j=1}^{N} \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j} \right) (x_j^0 - x_j), \qquad (18.1.8.1)$$

where the summation is over the N parameters of the model. This is simply the first-order expansion of $f_i(\mathbf{x})$ and expresses the hypothesis that the calculated values should match the observed values. The system of simultaneous *observational equations* can be solved for the parameter shifts provided that there are at least as many observations as there are parameters to be determined. When the number of observational equations exceeds the number of parameters, the least-squares solution is that which minimizes (18.1.4.1). This is the method generally used for refining small-molecule crystal structures, and increasingly for macromolecular structures at atomic resolution.

18.1.8.2. Normal equations

In matrix form, the observational equations are written as

$$\mathbf{A}\Delta = \mathbf{r},$$

where $\mathbf{A}$ is the M by N matrix of derivatives, Δ is the parameter shifts and $\mathbf{r}$ is the vector of residuals given on the left-hand sides of equation (18.1.8.1). The *normal equations* are formed by multiplying both sides of the equation by $\mathbf{A}^T$. This produces an N by N square system, the solution to which is the desired least-squares solution for the parameter shifts.

$$\mathbf{A}^T\mathbf{A}\Delta = \mathbf{A}^T\mathbf{r} \text{ or } \mathbf{M}\Delta = \mathbf{b},$$

$$m_{ij} = \sum_{k=1}^{M} w_k \left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right)\left(\frac{\partial f_k(\mathbf{x})}{\partial x_j}\right),$$

$$b_i = \sum_{k=1}^{M} w_k[y_k - f_k(\mathbf{x})]\left(\frac{\partial f_k(\mathbf{x})}{\partial x_i}\right).$$

Similar equations are obtained by expanding (18.1.4.1) as a second-order Taylor series about the minimum $\mathbf{x}_0$ and differentiating.

$$\Phi(\mathbf{x}-\mathbf{x}_0) \approx \Phi(\mathbf{x}_0) + \left\langle \left(\frac{\partial\Phi}{\partial x_i}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle + \frac{1}{2}\left\langle (\mathbf{x}-\mathbf{x}_0) \middle| \left(\frac{\partial^2\Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0} \middle| (\mathbf{x}-\mathbf{x}_0) \right\rangle,$$

$$\left|\left(\frac{\partial\Phi}{\partial\mathbf{x}}\right)\right\rangle \approx \left|\left(\frac{\partial^2\Phi}{\partial x_i \partial x_j}\right)_{\mathbf{x}_0}\right| \left.(\mathbf{x}-\mathbf{x}_0)\right\rangle.$$

The second-order approximation is equivalent to assuming that the matrix of second derivatives does not change and hence can be computed at $\mathbf{x}$ instead of at $\mathbf{x}_0$.

18.1.8.3. Choice of optimization method

First-order methods are generally the most economical for macromolecular problems. The most general approach is to treat the problem as a non-linear optimization problem from the beginning. This strategy is used by *TNT* (Tronrud *et al.*, 1987; Tronrud, 1997) and by *X-PLOR* (Kuriyan *et al.*, 1989), although by very different methods.

TNT uses a preconditioned conjugate gradient procedure (Tronrud, 1992), where the preconditioning function is the second derivatives of the objective function with respect to each parameter. In other words, at each step the parameters are normalized by the curvature with respect to that parameter, and a normal conjugate gradient step is taken. This has the effect that stiff parameters, which have steep derivatives, are scaled down, while soft parameters (such as B factors), which have small derivatives, are scaled up. This greatly increases both the rate and radius of convergence of the method.

X-PLOR (and its intellectual descendent, *CNS*) (Chapters 18.2 and 18.6) uses a simulated annealing procedure that derives sampling points by molecular dynamics. Simulated annealing is a process by which the objective function is sampled at a new point in parameter space. If the value of the objective function at the new point is less than that at the current point, the new point becomes the current point. If the value of the objective function is greater at the new point than at the current point, the Boltzmann probability $\exp(-\Delta E/kT)$ of the difference in function values ΔE is compared to a random number. If it is less than the random number, the new point is accepted as the current point; otherwise it is rejected. This process continues until a sufficiently deep minimum is found that the sampling process never leaves that region of parameter space. At this point the 'temperature' in the Boltzmann factor is reduced, which lowers the probability that the current point will move out of the region. This produces a finer search of the local region. The cooling process is continued until the solution has been restricted to a sufficiently small region. There are many variations of the strategy that affect the rate of convergence and the completeness of sampling. The primary virtue of simulated annealing is that it does not become trapped in shallow local minima. Simulated annealing can be either a zero-order method or a first-order method, depending on the strategy used to generate new sampling points. *X-PLOR* treats the fit to the diffraction data as an additional energy term, and the gradient of that 'energy' is treated as a force. This makes it a first-order method.

The first widely available macromolecular refinement program, *PROLSQ* (Konnert, 1976), uses an approximation to the second-order problem in which the matrix is greatly simplified. The parameters for each atom are treated as a small block on the diagonal of the matrix, and the off-diagonal blocks for pairs of atoms related by geometric constraints are also filled in. The sparse set of linear equations is then solved by an adaptation of the method of conjugate gradients.

The most comprehensive refinement program available for macromolecules is the same as the most comprehensive program available for small molecules – *SHELXL98* (Sheldrick, 1993; see also Chapter 18.9). The primary adaptations to macromolecular problems have been the addition of conjugate gradients as an optimization method for cases in which the full matrix will not fit in the available memory and facilities to process the polymeric models required for macromolecules.

18.1.8.4. Singularity in refinement

Unless there are more linearly independent observations than there are parameters to fit them, the system of normal equations has no solution. The inverse of the matrix does not exist. Second-order methods fail in these circumstances by doing the matrix equivalent of dividing by zero. However, the objective function is still defined and has a defined gradient at all points. First-order methods will find a point at which the gradient is close to zero, and zero-order methods will still find a minimum value for the objective function. The difficulty is that the points so found are not unique. If one computes the eigenvalues and eigenvectors of the matrix of normal equations, one will find in this case that there are some eigenvalues that are very small or zero. The eigenvectors corresponding to these eigenvalues define sets of directions in which the parameters can be moved without affecting the value of the objective function. This region of the parameter space simply cannot be determined by the available data. The only recourses are to modify the model so that it has fewer parameters, add additional restraints to the problem, or collect more data. The real hazard with this situation is that the commonly used refinement methods do not detect the problem. Careful use of cross validation and keeping careful count of the parameters are the only remedy.

18.1.9. Evaluation of the model

Macromolecular model refinement is a cyclic process. No presently known refinement algorithm can remove all the errors of chain tracing, conformation, or misinterpretation of electron density. Other methods must be interspersed with refinement to help remove model errors. These errors are detected by basic sanity checks and the use of common sense about the model. This topic is discussed comprehensively in Part 21 and in Kleywegt (2000).

18.1.9.1. Examination of outliers in the model

Refinement-program output listings will normally provide some information on atoms that are showing non-standard bond lengths, bond angles or B factors. In addition, there is other software which can help identify non-standard or unusual geometry, such as *PROCHECK* (Laskowski *et al.*, 1993) and *WHAT IF* (Vriend, 1990). These are very useful in identifying questionable regions of structure but should not be completely relied on to identify errors or how the molecular models may be improved. Overall, the constraints in the model must be satisfied exactly, and the restraints should have a statistically reasonable distribution of deviations from the ideal values.

18.1.9.2. Examination of model electron density

Refinement of the model to improve the agreement between the observed and calculated diffraction data and the associated calculated phases should result in improved electron-density and ΔF maps. Unexplained features in the electron-density map or difference map are a clear indication that the model is not yet complete or accurate. Careful examination of the Fourier maps is essential. Interactive graphics programs such as *XtalView* (McRee, 1993) and *O* contain a number of analysis tools to aid in the identification of errors in the models.

There are several different types of Fourier maps that can be useful in the correction of the models. This topic is discussed extensively in Chapter 15.2. Usual maps include F_o maps, ΔF maps and $(nF_o - mF_c)$ maps. The Fourier coefficients used to compute the maps should be weighted by estimates of the degree of bias as described in Chapter 15.2. While ΔF maps are very useful in highlighting areas in the maps that reflect the greatest difference between the F_o's and F_c's in Fourier space, they do not show the electron density of the unit cell. Positive and negative regions of a ΔF map may be the result of positional errors of an atom or group of atoms, B-factor errors, completely misplaced atoms or missing atoms. F_o maps show the electron density but are biased by the current model. A $(2F_o - F_c)$ map is a combination of an F_o map and a ΔF map which results in a map better showing the changes due to errors. Some investigators prefer using further amplified ΔF contributions by using a $(3F_o - 2F_c)$ map or higher-order terms.

The contribution of the disordered solvent continuum has been discussed previously. Macromolecular crystals also contain significant quantities of discrete or partially discrete solvent molecules (*i.e.* water). Care needs to be taken in adding solvent to a model. Errors in models generate peaks in Fourier maps that can be interpreted as solvent peaks. Hence, adding solvent peaks too early in the refinement process may, in fact, lead to model errors. Automatic water-adding programs are becoming more common; examples include *SHELXL98* and *ARP/wARP* (Lamzin & Wilson, 1997). These programs check if the waters are with in reasonable bonding distances of hydrogen-bonding atoms. There is a distribution of solvent molecules ranging from ones with low B factors at unit occupancy to ones with very large B factors. Various criteria are used to decide on a cutoff in the discrete solvent contribution. A rule of thumb for ambient-temperature data sets is frequently about one solvent molecule per residue in a protein molecule. As more data are being collected at cryogenic temperatures, this ratio is tending to go up. Noise is being fitted if too many peaks in a ΔF map are being assigned as solvent molecules. This can also contribute to reducing R factors on incorrect models. Solvent sites may not be fully occupied. Because of the large B factors and limited range of the diffraction data, the B factors and occupancy are highly correlated. Refinement of occupancy does not usually contribute either to improving a model or to reduction of R factors in structures with up to 2.0 Å resolution data. Beyond 1.5 Å data, it may be possible to refine solvent water occupancies and B factors. At even higher resolution, some programs, such as *SHELXL98*, provide anisotropic refinement methods which may further improve the solvent model while reducing R factors including $R_{\rm free}$.

18.1.9.3. R and R_{free}

Cross validation is a powerful tool for avoiding over-interpretation of the data by a too elaborate model. The introduction of cross validation to crystallography (Brünger, 1992) has been responsible for significant improvement in the quality of structure determinations. A subset of the reflections, chosen randomly, is segregated and not used in the refinement. If the model is correct and the only errors are statistical, these reflections should have an R factor close to that of the reflections used in the refinement. Changes to the model should affect both R and $R_{\rm free}$ similarly. Kleywegt & Jones (1997) have pointed out that it is necessary to treat the selection of free reflections very carefully in the presence of noncrystallographic symmetry.

18.1.10. Conclusion

It is always important to bear in mind that macromolecular crystal structures are models intended to explain a particular set of observations. Statistical measures can determine how well the model explains the observations, but cannot say whether the model is true or not. The distinction between precision and accuracy must always be kept in mind. The objective should not be simply to obtain the best fit of a model to the data, but, in addition, to find all of the ways in which a model does *not* fit the data and correct them. Until the day when all crystals diffract to atomic resolution, the primary objective of refinement of the models will be to determine just how well the structures are or are not determined.

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brunger, A. T. (1999). *Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. Acta Cryst.* D**55**, 181–190.

Booth, A. D. (1946*a*). *A differential Fourier method for refining atomic parameters in crystal structure analysis. Trans. Faraday Soc.* **42**, 444–448.

Booth, A. D. (1946*b*). *The simultaneous differential refinement of co-ordinates and phase angles in X-ray Fourier synthesis. Trans. Faraday Soc.* **42**, 617–619.

Bricogne, G. (1997). *Bayesian statistical viewpoint on structure determination: basic concepts and examples. Methods Enzymol.* **276**, 361–423.

Brünger, A. T. (1992). *Free R-value – a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Busing, W. R., Martin, K. O. & Levy, H. A. (1962). *ORFLS*. Report ORNL-TM-305. Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA.

Diamond, R. (1971). *A real-space refinement procedure for proteins*. *Acta Cryst.* A**27**, 436–452.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures*. *Methods Enzymol.* **115**, 252–270.

Hughes, E. W. (1941). *The crystal structure of melamine. J Am. Chem. Soc.* **63**, 1737–1752.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

International Tables for Crystallography (2008). Vol. B. *Reciprocal Space*, edited by U. Shmueli. Heidelberg: Springer.

Kleywegt, G. J. (2000). *Validation of protein crystal structures*. *Acta Cryst.* D**56**, 249–265.

Kleywegt, G. J. & Jones, T. A. (1997). *Model building and refinement practice. Methods Enzymol.* **277**, 208–230.

Konnert, J. H. (1976). *A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units*. *Acta Cryst.* A**32**, 614–617.

Konnert, J. H. & Hendrickson, W. A. (1980). *A restrained-parameter thermal-factor refinement procedure*. *Acta Cryst.* A**36**, 344–350.

Kuriyan, J., Brünger, A. T., Karplus, M. & Hendrickson, W. A. (1989). *X-ray refinement of protein structures by simulated annealing: test of the method on myohemerythrin*. *Acta Cryst.* A**45**, 396–409.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography*. In *Macromolecular Crystallography*, Part B, edited by C. C. & R. Sweet, 269–305. San Diego: Academic Press.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures*. *J. Appl. Cryst.* **26**, 283–291.

McRee, D. E. (1993). *Practical Protein Crystallography*, p. 386. San Diego: Academic Press.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method*. *Acta Cryst.* D**53**, 240–253.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Efficient anisotropic refinement of macromolecular structures using FFT*. *Acta Cryst.* D**55**, 247–255.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement*. *Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. *Acta Cryst.* A**52**, 659–668.

Prince, E. (1994). *Mathematical Techniques in Crystallography and Materials Science*. 2nd ed. Berlin: Springer-Verlag.

Rice, L. M. & Brünger, A. T. (1994). *Torsion-angle dynamics – reduced variable conformational sampling enhances crystallographic structure refinement*. *Proteins Struct. Funct. Genet.* **19**, 277–290.

Sheldrick, G. M. (1993). *SHELXL93. Program for the Refinement of Crystal Structures*. University of Göttingen, Germany.

Tronrud, D. E. (1992). *Conjugate-direction minimization: an improved method for the refinement of macromolecules*. *Acta Cryst.* **48**, 912–916.

Tronrud, D. E. (1996). *Knowledge-based B-factor restraints for the refinement of proteins*. *J. Appl. Cryst.* **29**, 100–104.

Tronrud, D. E. (1997). *The TNT refinement package*. *Methods Enzymol.* **277**, 306–318.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *An efficient general-purpose least-squares refinement program for macromolecular structures*. *Acta Cryst.* A**43**, 489–501.

Vriend, G. (1990). *WHAT IF: a molecular modeling and drug design program*. *J. Mol. Graphics*, **8**, 52–56.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1972). *The structure of a non-heme iron protein: rubredoxin at 1.5 Å resolution*. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 359–367.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Refinement of the model of a protein: rubredoxin at 1.5 Å resolution*. *Acta Cryst.* B**29**, 943–956.

International Tables for Crystallography (2012). Vol. F, Chapter 18.2, pp. 466–473.

Chapter 18.2. Enhanced macromolecular refinement by simulated annealing

A. T. Brunger, P. D. Adams and L. M. Rice

18.2.1. Introduction

The analysis of X-ray diffraction data generally requires sophisticated computational procedures that culminate in refinement and structure validation. The refinement procedure can be formulated as the chemically constrained or restrained nonlinear optimization of a target function, which usually measures the agreement between observed diffraction data and data computed from an atomic model. The ultimate goal of refinement is to optimize simultaneously the agreement of an atomic model with observed diffraction data and with *a priori* chemical information.

The target function used for this optimization normally depends on several atomic parameters and, most importantly, on atomic coordinates. The large number of adjustable parameters (typically at least three times the number of atoms in the model) gives rise to a very complicated target function. This, in turn, produces what is known as the multiple minima problem: the target function contains many local minima in addition to the global minimum, and this tends to defeat gradient-descent optimization techniques such as conjugate gradient or least-squares methods (Press *et al.*, 1986). These methods are unable to sample molecular conformations thoroughly enough to find the optimal model if the starting one is far from the correct structure.

The challenges of crystallographic refinement arise not only from the high dimensionality of the parameter space, but also from the phase problem. For new crystal structures, initial electron-density maps must be computed from a combination of observed diffraction amplitudes and experimental phases, where the latter are typically of poorer quality and/or at a lower resolution than the former. A different problem arises when structures are solved by molecular replacement (Hoppe, 1957; Rossmann & Blow, 1962), which uses a similar structure as a search model to calculate initial phases. In this case, the resulting electron-density maps can be severely 'model-biased', that is, they sometimes seem to confirm the existence of the search model without providing clear evidence of actual differences between it and the true crystal structure. In both cases, initial atomic models usually contain significant errors and require extensive refinement.

Simulated annealing (Kirkpatrick *et al.*, 1983) is an optimization technique particularly well suited to overcoming the multiple minima problem. Unlike gradient-descent methods, simulated annealing can cross barriers between minima and, thus, can explore a greater volume of the parameter space to find better models (deeper minima). Following its introduction to crystallographic refinement (Brünger *et al.*, 1987), there have been major improvements of the original method in four principal areas: the measure of model quality, the search of the parameter space, the target function and the modelling of conformational variability.

For crystallographic refinement, the introduction of cross validation and the free *R* value (Brünger, 1992) has significantly reduced the danger of overfitting the diffraction data during refinement. Cross validation also produces more realistic coordinate-error estimates based on the Luzzati or σ_A methods (Kleywegt & Brünger, 1996). The complexity of the conformational space has been reduced by the introduction of torsion-angle refinement methods (Diamond, 1971; Rice & Brünger, 1994), which decrease the number of adjustable parameters that describe a model approximately tenfold. The target function has been improved by using a maximum-likelihood approach which takes into account model error, model incompleteness and errors in the experimental data (Bricogne, 1991; Pannu & Read, 1996). Cross validation of parameters for the maximum-likelihood target function was essential in order to obtain better results than with conventional target functions (Pannu & Read, 1996; Adams *et al.*, 1997; Read, 1997). Finally, the sampling power of simulated annealing has been used for exploring the molecule's conformational space in cases where the molecule undergoes dynamic motion or exhibits static disorder (Kuriyan *et al.*, 1991; Burling & Brünger, 1994; Burling *et al.*, 1996).

18.2.2. Cross validation

Cross validation (Brünger, 1992) plays a fundamental role in the maximum-likelihood target functions described below. A few remarks about this method are therefore warranted (for reviews see Kleywegt & Brünger, 1996; Brünger, 1997). For cross validation, the diffraction data are divided into two sets: a large *working* set (usually comprising 90% of the data) and a complementary *test* set (comprising the remaining 10%). The diffraction data in the working set are used in the normal crystallographic refinement process, whereas the test data are not. The cross-validated (or 'free') *R* value computed with the test-set data is a more faithful indicator of model quality. It provides a more objective guide during the model building and refinement process than the conventional *R* value. It also ensures that introduction of additional parameters (*e.g.* water molecules, relaxation of noncrystallographic symmetry restraints, or multiconformer models) improves the quality of the model, rather than increasing overfitting.

Since the conventional *R* value shows little correlation with the accuracy of a model, coordinate-error estimates derived from the Luzzati (1952) or σ_A (Read, 1986) methods are unrealistically low. Kleywegt & Brünger (1996) showed that more reliable coordinate errors can be obtained by cross validation of the Luzzati or σ_A coordinate-error estimates. An example is shown in Fig. 18.2.2.1 using the crystal structure and diffraction data of penicillopepsin (Hsu *et al.*, 1977). At 1.8 Å resolution, the model has an estimated coordinate error of ∼0.2 Å as assessed by multiple independent refinements. As the resolution of the diffraction data is artificially truncated and the model re-refined, the coordinate error (assessed by the atomic root-mean-square difference to the refined model at 1.8 Å resolution) increases monotonically. The conventional *R* value improves as the resolution decreases and the quality of the model worsens. Consequently, coordinate-error estimates do not display the correct behaviour either: the error estimates are approximately constant, regardless of the resolution and actual coordinate error of the models. However, when cross validation is used (*i.e.*, the test reflections are used to compute the estimated coordinate errors),

Copyright © 2012 International Union of Crystallography

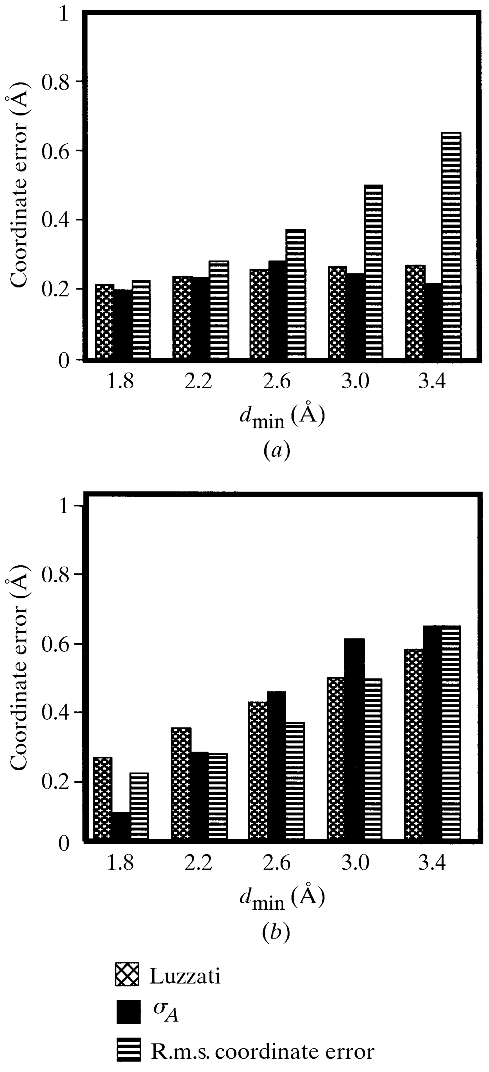


Figure 18.2.2.1

Effect of resolution on coordinate-error estimates: accuracy as a function of resolution. Refinements were begun with the crystal structure of penicillopepsin (Hsu *et al.*, 1977) with water molecules omitted and with uniform temperature factors. The low-resolution limit was set to 6 Å. Inclusion of all low-resolution diffraction data does not change the conclusions (Adams *et al.*, 1997). The penicillopepsin diffraction data were artificially truncated to the specified high-resolution limit. Each refinement consisted of simulated annealing using a Cartesian-space slow-cooling protocol starting at 2000 K, overall B-factor refinement and individual restrained B-factor refinement. All refinements were carried out with 10% of the diffraction data randomly omitted for cross validation. (*a*) Coordinate-error estimates of the refined structures using the methods of Luzzati (1952) and Read (1986). All observed diffraction data were used, *i.e.* no cross validation was performed. The actual coordinate errors (r.m.s. differences to the original crystal structure) are shown for comparison. (*b*) Cross-validated coordinate-error estimates. The test set was used to compute the coordinate-error estimates (Kleywegt & Brünger, 1996).

the results are much better: the cross-validated errors are close to the actual coordinate error, and they show the correct trend as a function of resolution (Fig. 18.2.2.1).

18.2.3. The target function

Crystallographic refinement is a search for the global minimum of the target

$$E = E_{\text{chem}} + w_{\text{X-ray}} E_{\text{X-ray}} \qquad (18.2.3.1)$$

as a function of the parameters of an atomic model, in particular, atomic coordinates. E_{chem} comprises empirical information about chemical interactions; it is a function of all atomic positions, describing covalent (bond lengths, bond angles, torsion angles, chiral centres and planarity of aromatic rings) and non-bonded (intramolecular as well as intermolecular and symmetry-related) interactions (Hendrickson, 1985). $E_{\text{X-ray}}$ is related to the difference between observed and calculated data, and $w_{\text{X-ray}}$ is a weight appropriately chosen to balance the gradients (with respect to atomic parameters) arising from the two terms.

18.2.3.1. X-ray diffraction data versus model

The traditional form of $E_{\text{X-ray}}$ consists of the crystallographic residual, E^{LSQ}, defined as the sum over the squared differences between the observed ($|\mathbf{F}_o|$) and calculated ($|\mathbf{F}_c|$) structure-factor amplitudes for a particular atomic model:

$$E_{\text{X-ray}} = E^{\text{LSQ}} = \sum_{hkl\in\text{working set}} (|\mathbf{F}_o| - k|\mathbf{F}_c|)^2, \qquad (18.2.3.2)$$

where hkl are the indices of the reciprocal-lattice points of the crystal and k is a relative scale factor.

Minimization of E^{LSQ} can produce improvement in the atomic model, but it can also accumulate systematic errors in the model by fitting noise in the diffraction data (Silva & Rossmann, 1985). The least-squares residual is a limiting case of the more general maximum-likelihood theory and is only justified if the model is nearly complete and error-free. These assumptions may be violated during the initial stages of refinement. Improved targets for macromolecular refinement have been obtained using the more general maximum-likelihood formulation (Bricogne, 1991; Pannu & Read, 1996; Adams *et al.*, 1997; Murshudov *et al.*, 1997). The goal of the maximum-likelihood method is to determine the likelihood of the model, given estimates of the model's errors and those of the measured intensities.

A starting point for the maximum-likelihood formulation of crystallographic refinement is the Sim (1959) distribution, *i.e.*, the Gaussian conditional probability distribution of the 'true' structure factors, **F**, given a partial model with structure factors $\mathbf{F}_c$ and the model's error (Fig. 18.2.3.1) (Srinivasan, 1966; Read, 1986, 1990) (for simplicity we will only discuss the case of acentric reflections),

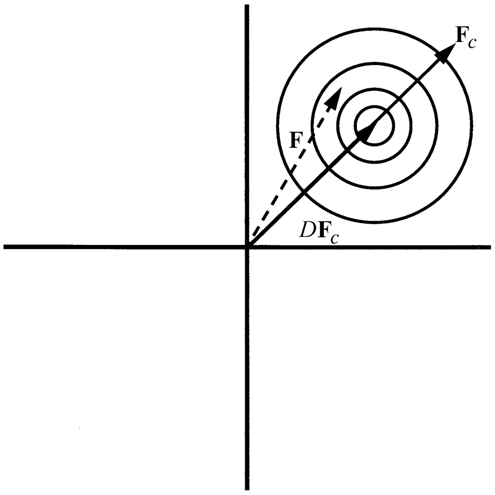


Figure 18.2.3.1

The Gaussian probability distribution forms the basis of maximum-likelihood targets in crystallographic refinement. The conditional probability of the true structure factor, **F**, given model structure factors, is a Gaussian in the complex plane [equation (18.2.3.3)]. The expected value of the probability distribution is $D\mathbf{F}_c$ with variance σ_Δ, where D and σ_Δ account for missing or incorrectly placed atoms in the model.

$$P_a(\mathbf{F};\mathbf{F}_c) = (1/\pi\varepsilon\sigma_\Delta^2)\exp[-(\mathbf{F} - D\mathbf{F}_c)^2/\varepsilon\sigma_\Delta^2], \qquad (18.2.3.3)$$

where σ_Δ is a parameter that incorporates the effect of the fraction of the asymmetric unit that is missing from the model and errors in the partial structure. Assuming a Wilson distribution of intensities, it can be shown that (Read, 1990)

$$\sigma_\Delta^2 = \langle|\mathbf{F}_o|^2\rangle - D^2\langle|\mathbf{F}_c|^2\rangle, \qquad (18.2.3.4)$$

where D is a factor that takes into account model error: it is unity in the limiting case of an error-free model and it is zero if no model is available (Luzzati, 1952; Read, 1986). For a complete and error-free model, σ_Δ therefore becomes zero, and the probability distribution, $P_a(\mathbf{F};\mathbf{F}_c)$, is infinitely sharp.

Taking measurement errors into account requires multiplication of equation (18.2.3.3) with an appropriate probability distribution (usually a conditional Gaussian distribution with standard deviation σ_o) of the observed structure-factor amplitudes ($|\mathbf{F}_o|$) around the 'true' structure-factor amplitudes ($|\mathbf{F}|$),

$$P_{\rm meas}(|\mathbf{F}_o|;|\mathbf{F}|). \qquad (18.2.3.5)$$

Prior knowledge of the phases of the structure factors can be incorporated by multiplying equation (18.2.3.3) with a phase probability distribution

$$P_{\rm phase}(\varphi) \qquad (18.2.3.6)$$

and rewriting equation (18.2.3.3) in terms of the structure-factor moduli and amplitudes of $\mathbf{F} = |\mathbf{F}|\exp(i\varphi)$.

The unknown variables $|\mathbf{F}|$ and φ in equations (18.2.3.3)–(18.2.3.5) have to be eliminated by integration in order to obtain the conditional probability distribution of the observed structure-factor amplitudes, given a partial model with errors, the amplitude measurement errors and prior phase information:

$$\begin{aligned}P_a(|\mathbf{F}_o|;\mathbf{F}_c) &= (1/\pi\varepsilon\sigma_\Delta^2)\textstyle\int {\rm d}\varphi\,{\rm d}|\mathbf{F}|;\ |\mathbf{F}|P_{\rm meas}(|\mathbf{F}_o|;|\mathbf{F}|)\\ &\quad\times P_{\rm phase}(\varphi)\exp\left\{-[|\mathbf{F}|\exp(i\varphi) - D\mathbf{F}_o]^2/\varepsilon\sigma_\Delta^2\right\}.\end{aligned} \qquad (18.2.3.7)$$

The likelihood, $\mathcal{L}$, of the model is defined as the joint probability distribution of the structure factors of all reflections in the working set. Assuming independent and uncorrelated structure factors, $\mathcal{L}$ is simply the product of the distributions in equation (18.2.3.7) for all reflections. Instead of maximizing the likelihood, it is more common to minimize the negative logarithm of the likelihood,

$$E_{\text{X-ray}} = \mathcal{L} = -\sum_{hkl\in\text{working set}}\log[P_a(|\mathbf{F}_o|;\mathbf{F}_c)]. \qquad (18.2.3.8)$$

Empirical estimates of σ_Δ [and D through equation (18.2.3.4)] can be obtained by minimizing $\mathcal{L}$ for a particular atomic model. It is generally assumed that σ_Δ and D show relatively little variation among neighbouring reflections. Accepting this assumption, σ_Δ and D can be estimated by considering narrow resolution shells of reflections and assuming that the two parameters are constant in these shells. Minimization of $\mathcal{L}$ can then be performed as a function of these constant shell parameters while keeping the atomic model fixed (Read, 1986, 1997). Alternatively, one can assume a two-term Gaussian model for σ_Δ (Murshudov *et al.*, 1997) and minimize $\mathcal{L}$ as a function of the Gaussian parameters. Note that individual atomic B factors are taken into account by the calculated model structure factors ($\mathbf{F}_c$).

This empirical approach to estimate σ_Δ and D requires occasional recomputation of these values as the model improves. Refinement methods that improve the model structure factors, $\mathbf{F}_c$, will therefore have a beneficial effect on σ_Δ and D. Better estimates of these values will then enhance the next refinement cycle. Thus, powerful optimization methods and maximum-likelihood targets are expected to interact in a synergistic fashion (*cf.* Fig. 18.2.5.1). Structure-factor averaging of multi-start refinement models can provide another layer of improvement by producing a better description of $\mathbf{F}_c$ if the model shows significant variability due to errors or intrinsic flexibility (see below).

In order to achieve an improvement over the least-squares residual [equation (18.2.3.2)], cross validation was found to be essential (Pannu & Read, 1996; Adams *et al.*, 1997) for the estimation of model incompleteness and errors (σ_Δ and D). Since the test set typically contains only 10% of the diffraction data, these cross-validated quantities can show significant statistical fluctuations as a function of resolution. In order to reduce these fluctuations, Read (1997) devised a smoothing method by applying restraints to σ_A values between neighbouring resolution shells where

$$\sigma_A = \left[1 - (\sigma_\Delta/\langle|\mathbf{F}_o|^2\rangle)\right]^{1/2}. \qquad (18.2.3.9)$$

Pannu & Read (1996) have developed an efficient Gaussian approximation of equation (18.2.3.7) in cases of no prior phase information, termed the 'MLF' target function. In the limit of a perfect model (*i.e.* $\sigma_\Delta = 0$ and $D = 1$), MLF reduces to the traditional least-squares residual [equation (18.2.3.2)] with $1/\sigma_o^2$ weighting. In the case of prior phase information, the integration over the phase angles has been carried out numerically in equation (18.2.3.7), termed the 'MLHL' target (Pannu *et al.*, 1998). A maximum-likelihood function which expresses equation (18.2.3.7) in terms of observed intensities has also been developed, termed 'MLI' (Pannu & Read, 1996).

18.2.3.2. *A priori chemical information*

The parameters for the covalent terms in $E_{\rm chem}$ [equation (18.2.3.1)] can be derived from the average geometry and (r.m.s.) deviations observed in a small-molecule database. Extensive statistical analyses were undertaken for the chemical moieties of proteins (Engh & Huber, 1991) and polynucleotides (Parkinson *et al.*, 1996) using the Cambridge Structural Database (Allen *et al.*, 1983). Analysis of the ever-increasing number of atomic resolution macromolecular crystal structures will no doubt cause some modifications of these parameters in the future.

It is common to use a purely repulsive quartic function ($E_{\rm repulsive}$) for the non-bonded interactions that are included in $E_{\rm chem}$ (Hendrickson, 1985):

$$E_{\rm repulsive} = \sum_{ij}[(cR_{ij}^{\min})^n - R_{ij}^n]^m, \qquad (18.2.3.10)$$

where R_{ij} is the distance between two atoms i and j, $R_{ij}^{\min}$ is the van der Waals radius for a particular atom pair ij, $c \leq 1$ is a constant that is sometimes used to reduce the radii, and $n = 2$, $m = 2$ or $n = 1$, $m = 4$. van der Waals attraction and electrostatic interactions are usually not included in crystallographic refinement. These simplifications are valid since the diffraction data contain information that is able to produce atomic conformations consistent with actual non-bonded interactions. In fact, atomic resolution crystal structures can be used to derive parameters for electrostatic charge distributions (Pearlman & Kim, 1990).

18.2.4. Searching conformational space

Annealing denotes a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase and is then cooled slowly so that all particles arrange themselves in the lowest energy state. By formally defining the target, E [equation (18.2.3.1)], to be the equivalent of the potential energy of the system, one can simulate such an annealing process (Kirkpatrick *et al.*, 1983). There is no guarantee that simulated annealing will find the global minimum (Laarhoven & Aarts, 1987). However, compared to conjugate-gradient minimization, where search directions must follow the gradient, simulated annealing achieves more optimal solutions by allowing motion against the gradient (Kirkpatrick *et al.*, 1983). The likelihood of *uphill* motion is determined by a control parameter referred to as *temperature*. The higher the temperature, the more likely it is that simulated annealing will overcome barriers (Fig. 18.2.4.1). It should be noted that the simulated-annealing temperature normally has no physical meaning and merely determines the likelihood of overcoming barriers of the target function in equation (18.2.3.1).

The simulated-annealing algorithm requires a mechanism to create a Boltzmann distribution at a given temperature, T, and an annealing schedule, that is, a sequence of temperatures $T_1 \geq T_2 \geq \ldots \geq T_l$ at which the Boltzmann distribution is computed. Implementations differ in the way they generate a transition, or move, from one set of parameters to another that is consistent with the Boltzmann distribution at a given temperature. The two most widely used methods are Metropolis Monte Carlo (Metropolis *et al.*, 1953) and molecular dynamics (Verlet, 1967) simulations. For X-ray crystallographic refinement, molecular dynamics has proven extremely successful (Brünger *et al.*, 1987) because it limits the search to physically reasonable 'moves'.

18.2.4.1. Molecular dynamics

A suitably chosen set of atomic parameters can be viewed as generalized coordinates that are propagated in time by the classical equations of motion (Goldstein, 1980). If the generalized coordinates represent the x, y, z positions of the atoms of a molecule, the classical equations of motion reduce to the familiar Newton's second law:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = -\nabla_i E. \qquad (18.2.4.1)$$

The quantities m_i and $\mathbf{r}_i$ are, respectively, the mass and coordinates of atom i, and E is given by equation (18.2.3.1). The solution of the partial differential equations (18.2.4.1) can be achieved numerically using finite-difference methods (Verlet, 1967; Abramowitz & Stegun, 1968). This approach is referred to as molecular dynamics.

Initial velocities for the integration of equation (18.2.4.1) are usually assigned randomly from a Maxwell distribution at the appropriate temperature. Assignment of different initial velocities will generally produce a somewhat different structure after simulated annealing. By performing several refinements with different initial velocities, one can therefore improve the chances of success of simulated-annealing refinement. Furthermore, this improved sampling can be used to study discrete disorder and conformational variability, especially when using torsion-angle molecular dynamics (see below).

Although Cartesian (*i.e.* flexible bond lengths and bond angles) molecular dynamics places restraints on bond lengths and bond angles [through E_{chem}, equation (18.2.3.1)], one might want to implement these restrictions as constraints, *i.e.*, fixed bond lengths and bond angles (Diamond, 1971). This is supported by the observation that the deviations from ideal bond lengths and bond angles are usually small in macromolecular X-ray crystal structures. Indeed, fixed-length constraints have been applied to crystallographic refinement by least-squares minimization (Diamond, 1971). It is only recently, however, that efficient and robust algorithms have become available for molecular dynamics in torsion-angle space (Bae & Haug, 1987, 1988; Jain *et al.*, 1993; Rice & Brünger, 1994). We chose an approach that retains the Cartesian-coordinate formulation of the target function and its derivatives with respect to atomic coordinates, so that the calculation remains relatively straightforward and can be applied to any macromolecule or their complexes (Rice & Brünger, 1994). In this formulation, the expression for the acceleration becomes a function of positions and velocities. Iterative equations of motion for constrained dynamics in this formulation can be derived and solved by finite-difference methods (Abramowitz & Stegun, 1968). This method is numerically very robust and has a significantly increased radius of convergence in crystallographic refinement compared to Cartesian molecular dynamics (Rice & Brünger, 1994).

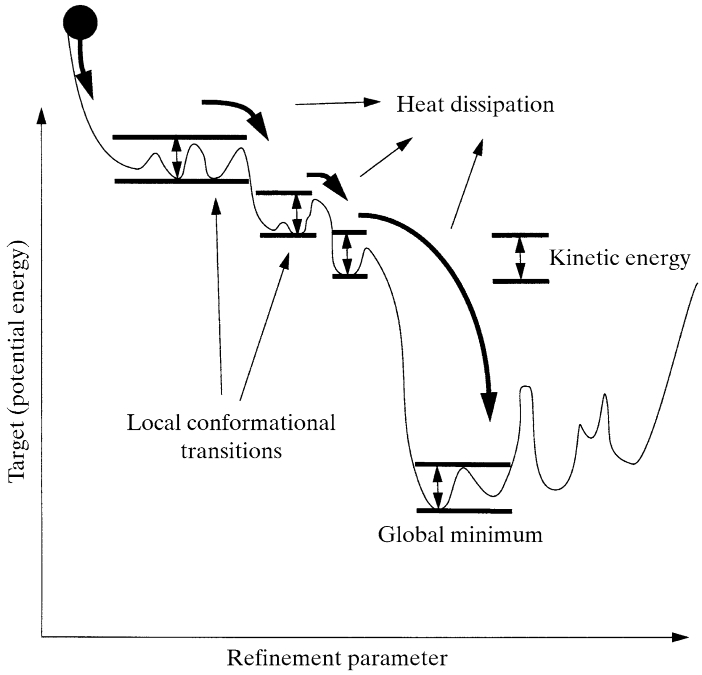


Figure 18.2.4.1
Illustration of simulated annealing for minimization of a one-dimensional function. The kinetic energy of the system (a 'ball' rolling on the one-dimensional surface) allows local conformational transitions with barriers smaller than the kinetic energy. If a larger drop in energy is encountered, the excess kinetic energy is dissipated. It is thus unlikely that the system can climb out of the global minimum once it has reached it.

18.2.4.2. Temperature control

Simulated annealing requires the control of the temperature during molecular dynamics. The current temperature of the simulation (T_{curr}) is computed from the kinetic energy

$$E_{\text{kin}} = \sum_i^n \tfrac{1}{2} m_i \left(\frac{\partial r_i}{\partial t} \right)^2 \qquad (18.2.4.2)$$

of the molecular-dynamics simulation,

$$T_{\text{curr}} = 2E_{\text{kin}}/3nk_B. \qquad (18.2.4.3)$$

Here, n is the number of atoms, m_i is the mass of the atom and k_B is Boltzmann's constant. One commonly used approach to control the temperature of the simulation consists of coupling the equations of motion to a heat bath through a 'friction' term (Berendsen *et al.*, 1984). Another approach is to rescale periodically the velocities in order to match $T_{\rm curr}$ with the target temperature.

18.2.4.3. Annealing schedules

The simulated-annealing temperature needs to be high enough to allow conformational transitions, but not so high that the model moves too far away from the correct structure. The optimal temperature for a given starting structure is a matter of trial and error. Starting temperatures that work for the average case have been determined for a variety of simulated-annealing protocols (Brünger, 1988; Adams *et al.*, 1997). However, it might be worth trying a different temperature if a particularly difficult refinement problem is encountered. In particular, significantly higher temperatures are attainable using torsion-angle molecular dynamics. Note that each simulated-annealing refinement is subject to 'chance' by using a random-number generator to generate the initial velocities. Thus, multiple simulated annealing runs can be carried out in order to increase the success rate of the refinement. The best structure(s) (as determined by the free *R* value) among a set of refinements using different initial velocities and/or temperatures can be taken for further refinement or structure-factor averaging (see below).

The annealing schedule can, in principle, be any function of the simulation step (or 'time' domain). The two most commonly used protocols are linear slow-cooling or constant-temperature followed by quenching. A slight advantage is obtained with slow cooling (Brünger *et al.*, 1990). The duration of the annealing schedule is another parameter. Too short a protocol does not allow sufficient sampling of conformational space. Too long a protocol may waste computer time, since it is more efficient to run multiple trials than one long refinement protocol (unpublished results).

18.2.4.4. An intuitive explanation of simulated annealing

The goal of any optimization problem is to find the global minimum of a target function. In the case of crystallographic refinement, one searches for the conformation or conformations of the molecule that best fit the diffraction data and that simultaneously maintain reasonable covalent and non-covalent interactions. Simulated-annealing refinement has a much larger radius of convergence than conjugate-gradient minimization (see below). It must, therefore, be able to find a lower minimum of the target *E* [equation (18.2.3.1)] than the local minimum found by simply moving along the negative gradient of *E*.

It is most easy to visualize this property of simulated annealing in the case of a one-dimensional problem, where the goal is to find the global minimum of a function with multiple minima (Fig. 18.2.4.1). An intuitive way to understand a molecular-dynamics simulation is to envisage a ball rolling on this one-dimensional surface. When the ball is far from the global minimum, it gains a certain momentum which allows it to cross barriers of the target function [equation (18.2.4.3)]. Slow-cooling temperature control ensures that the ball will eventually reach the global minimum rather than just bouncing across the surface. The initial temperature must be large enough to overcome smaller barriers, but low enough to ensure that the system will not escape the global minimum if it manages to arrive there.

While temperature itself is a global parameter of the system, temperature fluctuations arise principally from local conformational transitions, for example, from an amino-acid side chain falling into the correct orientation. These local changes tend to lower the value of the target *E*, thus increasing the kinetic energy, and hence the temperature, of the system. Once the temperature control has removed this excess kinetic energy through 'heat dissipation', the reverse transition is very unlikely, since it would require a localized increase in kinetic energy where the conformational change occurred in the first place (Fig. 18.2.4.1). Temperature control maintains a sufficient amount of kinetic energy to allow local conformational corrections, but does not supply enough to allow escape from the global minimum. This explains the observation that, on average, the agreement with the diffraction data will improve, rather than worsen, with simulated annealing.

18.2.5. Examples

Many examples have shown that simulated-annealing refinement starting from initial models (obtained by standard crystallographic techniques) produces significantly better final models compared to those produced by least-squares or conjugate-gradient minimization (Brünger *et al.*, 1987; Brünger, 1988; Fujinaga *et al.*, 1989; Kuriyan *et al.*, 1989; Rice & Brünger, 1994; Adams *et al.*, 1997). In another realistic test case (Adams *et al.*, 1999), a series of models for the aspartic proteinase penicillopepsin were generated from homologous structures present in the Protein Data Bank. The sequence identity among these structures ranged from 100% to 25%, thus providing a set of models with increasing coordinate error compared to the refined structure of penicillopepsin. These models, after truncation of all

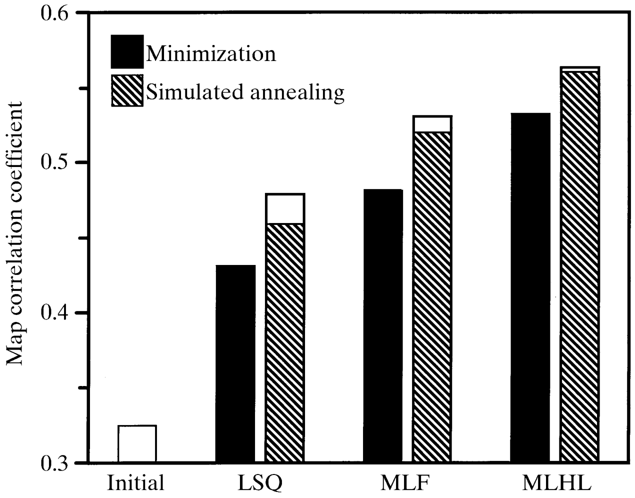


Figure 18.2.5.1
Simulated annealing produces better models than extensive conjugate-gradient minimization. Map correlation coefficients were computed before and after refinement against the native penicillopepsin diffraction data (Hsu *et al.*, 1977) for the polyalanine model derived from Rhizopuspepsin (Suguna *et al.*, 1987, PDB code 2APR). Correlation coefficients are between σ_A-weighted maps calculated from each model and from the published penicillopepsin structure. The observed penicillopepsin diffraction data were in space group *C*2 with cell dimensions $a = 97.37$, $b = 46.64$, $c = 65.47$ Å and $\beta = 115.4°$. All refinements were carried out using diffraction data from the lowest-resolution limit of 22.0 Å up to 2.0 Å. The MLHL refinements used single isomorphous phases from a $K_3UO_2F_5$ derivative of the penicillopepsin crystal structure, which covered a resolution range of 22.0 Å to 2.8 Å. Simulated-annealing refinements were repeated five times with different initial velocities. The numerical averages of the map correlation coefficients for the five refinements are shown as hashed bars. The best map correlation coefficients from simulated annealing are shown as white bars.

residues to alanine, were all used as search models in molecular replacement against the native penicillopepsin diffraction data. In all cases, the correct placement of the model in the penicillopepsin unit cell was found.

Both conjugate-gradient minimization and simulated annealing were carried out in order to compare the performance of the $E^{\rm LSQ}$ least-squares residual [equation (18.2.3.2)], MLF (the maximum-likelihood target using amplitudes) and MLHL (the maximum-likelihood target using amplitudes and experimental phase information). In the latter case, phases from single isomorphous replacement (SIR) were used. A very large number of conjugate-gradient cycles were carried out in order to make the computational requirements equivalent for both minimization and simulated annealing. The conjugate-gradient minimizations were converged, *i.e.* there was no change when further cycles were carried out.

For a given target function, simulated annealing always outperformed minimization (Fig. 18.2.5.1). For a given starting model, the maximum-likelihood targets outperformed the least-squares-residual target for both minimization and simulated annealing, producing models with lower phase errors and higher map correlation coefficients when compared with the published penicillopepsin crystal structure (Fig. 18.2.5.1). This improvement is illustrated in σ_A-weighted electron-density maps obtained from the resulting models (Fig. 18.2.5.2). The incorporation of experimental phase information further improved the refinement significantly despite the ambiguity in the SIR phase probability distributions. Thus, the most efficient refinement will make use of simulated annealing and phase information in the MLHL maximum-likelihood target function.

Cross validation is essential in the calculation of the maximum-likelihood target (Kleywegt & Brünger, 1996; Pannu & Read, 1996; Adams *et al.*, 1997). Maximum-likelihood refinement without cross validation gives much poorer results, as indicated by higher free R values, $R_{\rm free} - R$ differences and phase errors (Adams *et al.*, 1997). It should be noted that the final normal R value is in general increased, compared to refinements with the least-squares target, when using the cross-validated maximum-likelihood formulation. This is a consequence of the reduction of overfitting by this method.

18.2.6. Multi-start refinement and structure-factor averaging

Multiple simulated-annealing refinements starting from the same model, termed 'multi-start' refinement, will generally produce somewhat different structures. Even well refined structures will show some variation consistent with the estimated coordinate error of the model (*cf.* results for 1.8 Å resolution in Fig. 18.2.2.1). More importantly, the poorer the model, the more variation is observed (Brünger, 1988). Some of the models resulting from multi-start refinement may be better than others, for example, as judged by the free R value. Thus, if computer time is available, multi-start refinement has several advantages. A more optimal single model than that produced by a single simulated-annealing calculation can usually be obtained. Furthermore, each separate model coming from a multi-start refinement fits the data slightly differently. This could be the result of intrinsic flexibility in the molecule (see below) or the result of model-building error. Regions in the starting model that contain significant errors often show increased variability after multi-start refinement, and a visual inspection of the ensemble of models produced can be helpful in identifying these incorrectly modelled regions.

To better identify the correct conformation, structure factors from each of the models can be averaged (Rice *et al.*, 1998). This averaging tends to reduce the effect of local errors (noise) that are presumably different in each member of the family. The average structure factor can produce phases that contain less model bias than phases computed from a single model. It should also produce better estimates of σ_Δ and D for maximum-likelihood targets and for σ_A-weighted electron-density maps, because $\mathbf{F}_c$ is used in the computation of these parameters [equation (18.2.3.7)]. Because it is inherently a noise-reducing technique, multi-start refinement followed by structure-factor averaging should be most useful in situations where there is significant noise, namely when the data-to-parameter ratio is low (*e.g.* if only moderate-resolution diffraction data are available).

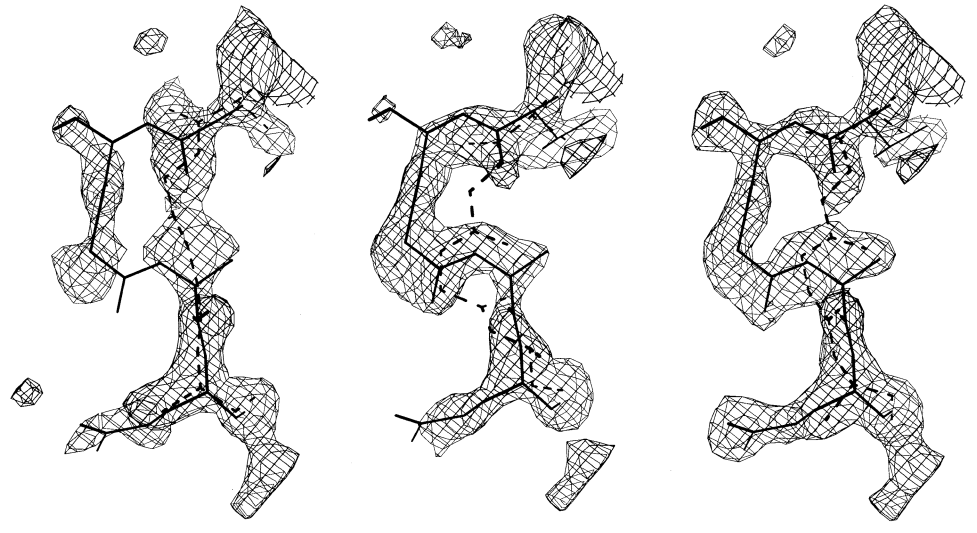


Figure 18.2.5.2
Maximum-likelihood targets significantly decrease model bias in simulated-annealing refinement. σ_A-weighted electron-density maps contoured at 1.25σ for models from simulated-annealing refinement with different targets are shown. Residues 233 to 237 are shown for the published penicillopepsin crystal structure (Hsu *et al.*, 1977) as solid lines, and for the model with the lowest free R value from five independent refinements as dashed lines.

18.2.7. Ensemble models

In cases of conformational variability or discrete disorder, there is not one single correct solution to the global minimization of equation (18.2.3.1). Rather, the X-ray diffraction data represent a spatial and temporal average over all conformations that are assumed by the molecule. Ensembles of structures, which are simultaneously refined against the observed data, may thus be a more appropriate description of the diffraction data. This has been used for some time when alternate conformations are modelled locally. Alternate conformations can be generalized to global conformations (Gros *et al.*, 1990; Kuriyan *et al.*, 1991; Burling & Brünger, 1994), *i.e.*, the model is duplicated n-fold, the calculated structure factors corresponding to each copy of the model are summed, and this composite structure factor is refined against the observed

X-ray diffraction data. Each member of the family is chemically 'invisible' to all other members. The optimal number, n, can be determined by cross validation (Burling & Brünger, 1994; Burling *et al.*, 1996).

An advantage of a multi-conformer model is that it directly incorporates many possible types of disorder and motion (global disorder, local side-chain disorder, local wagging and rocking motions). Furthermore, it can be used to detect automatically the most variable regions of the molecule by inspecting the atomic r.m.s. difference around the mean as a function of residue number. Thermal factors of single-conformer models may sometimes be misleading because they underestimate the degree of motion or disorder (Kuriyan *et al.*, 1986), and, thus, the multiple-conformer model can be a more faithful representation of the diffraction data. A disadvantage of the multi-conformer model is that it introduces many more parameters in the refinement.

Although there are some similarities between averaging structure factors of individually refined structures and performing multi-conformer refinement, there are also fundamental differences. For example, multi-start averaging seeks to improve the calculated electron-density map by averaging out the noise present in the individual models, each of which is still a good representation of the diffraction data. This method is most useful at the early stages of refinement when the model still contains errors. In contrast, multi-conformer refinement seeks to create an ensemble of structures at the final stages of refinement which, taken together, best represent the data. It should be noted that each individual conformer of the ensemble does not necessarily remain a good description of the diffraction data, since the whole ensemble is refined against the data. Clearly, multi-conformer refinement requires a high observable-to-parameter ratio.

18.2.8. Conclusions

Simulated annealing has dramatically improved the efficiency of crystallographic refinement. A case in point is the combination of torsion-angle molecular dynamics with cross-validated maximum-likelihood targets. These two independent developments interact synergistically to produce less model bias than any other method to date. The combined method dramatically increases the radius of convergence, allowing the productive refinement of poor initial models, *e.g.* those obtained by weak molecular-replacement solutions (Rice & Brünger, 1994; Adams *et al.*, 1997, 1999).

Simulated annealing can also be used to provide new physical insights into molecular function which may depend on conformational variability. The sampling characteristics of simulated annealing allow the generation of multi-conformer models that can represent molecular motion and discrete disorder, especially when combined with the acquisition of high-quality data (Burling *et al.*, 1996). Thus, simulated annealing is also a stepping stone towards development of improved models of macromolecules in solution and in the crystalline state.

The computational developments discussed in this review are implemented in the software suite *Crystallography & NMR System* (Brunger *et al.*, 1998). A pre-release of the software suite is available upon request.

LMR is an HHMI predoctoral fellow. This work was funded in part by grants from the National Science Foundation to ATB (BIR 9514819 and ASC 93-181159).

References

Abramowitz, M. & Stegun, I. (1968). *Handbook of Mathematical Functions. Applied Mathematics Series*, Vol. 55, p. 896. New York: Dover Publications.

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Adams, P. D., Pannu, N. S., Read, R. J. & Brunger, A. T. (1999). *Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. Acta Cryst.* D**55**, 181–190.

Allen, F. H., Kennard, O. & Taylor, R. (1983). *Systematic analysis of structural data as a research technique in organic chemistry. Acc. Chem. Res.* **16**, 146–153.

Bae, D.-S. & Haug, E. J. (1987). *A recursive formulation for constrained mechanical system dynamics: Part I. Open loop systems. Mech. Struct. Mach.* **15**, 359–382.

Bae, D.-S. & Haug, E. J. (1988). *A recursive formulation for constrained mechanical system dynamics: Part II. Closed loop systems. Mech. Struct. Mach.* **15**, 481–506.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). *Molecular dynamics with coupling to an external bath. J. Chem. Phys.* **81**, 3684–3690.

Bricogne, G. (1991). *A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data. Acta Cryst.* A**47**, 803–829.

Brunger, A. T. (1988). *Crystallographic refinement by simulated annealing: application to a 2.8 Å resolution structure of aspartate aminotransferase. J. Mol. Biol.* **203**, 803–816.

Brunger, A. T. (1992). *The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–474.

Brunger, A. T. (1997). *Free R value: cross-validation in crystallography. Methods Enzymol.* **277**, 366–396.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Cryst.* D**54**, 905–921.

Brunger, A. T., Krukowski, A. & Erickson, J. W. (1990). *Slow-cooling protocols for crystallographic refinement by simulated annealing. Acta Cryst.* A**46**, 585–593.

Brunger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Burling, F. T. & Brunger, A. T. (1994). *Thermal motion and conformational disorder in protein crystal structures: comparison of multi-conformer and time-averaging models. Isr. J. Chem.* **34**, 165–175.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brunger, A. T. (1996). *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. Science*, **271**, 72–77.

Diamond, R. (1971). *A real-space refinement procedure for proteins. Acta Cryst.* A**27**, 436–452.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray structure refinement. Acta Cryst.* A**47**, 392–400.

Fujinaga, M., Gros, P. & van Gunsteren, W. F. (1989). *Testing the method of crystallographic refinement using molecular dynamics. J. Appl. Cryst.* **22**, 1–8.

Goldstein, H. (1980). *Classical Mechanics*. 2nd ed. Reading, Massachusetts: Addison-Wesley.

Gros, P., van Gunsteren, W. F. & Hol, W. G. J. (1990). *Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science*, **249**, 1149–1152.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures. Methods Enzymol.* **115**, 252–270.

Hoppe, W. (1957). *Die 'Faltmolekülmethode' – eine neue Methode zur Bestimmung der Kristallstruktur bei ganz oder teilweise bekannter Molekülstruktur. Acta Cryst.* **10**, 750–751.

Hsu, I. N., Delbaere, L. T. J., James, M. N. G. & Hoffman, T. (1977). *Penicillopepsin from Penicillium janthinellum crystal structure at 2.8 Å and sequence homology with porcine pepsin. Nature (London)*, **266**, 140–145.

Jain, A., Vaidehi, N. & Rodriguez, G. (1993). *A fast recursive algorithm for molecular dynamics simulation. J. Comput. Phys*. **106**, 258–268.

Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Jr (1983). *Optimization by simulated annealing. Science*, **220**, 671–680.

Kleywegt, G. J. & Brunger, A. T. (1996). *Cross-validation in crystallography: practice and applications. Structure*, **4**, 897–904.

Kuriyan, J., Brunger, A. T., Karplus, M. & Hendrickson, W. A. (1989). *X-ray refinement of protein structures by simulated annealing: test of the method on myohemerythrin. Acta Cryst*. A**45**, 396–409.

Kuriyan, J., Ösapay, K., Burley, S. K., Brunger, A. T., Hendrickson, W. A. & Karplus, M. (1991). *Exploration of disorder in protein structures by X-ray restrained molecular dynamics. Proteins*, **10**, 340–358.

Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. (1986). *Effect of anisotropy and anharmonicity on protein crystallographic refinement. J. Mol. Biol*. **190**, 227–254.

Laarhoven, P. J. M. & Aarts, E. H. L. (1987). Editors. *Simulated Annealing: Theory and Applications*. Dordrecht: D. Reidel Publishing Company.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst*. **5**, 802–810.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A. & Teller, E. (1953). *Equation of state calculations by fast computing machines. J. Chem. Phys*. **21**, 1087–1092.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst*. D**53**, 240–255.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst*. D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst*. A**52**, 659–668.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brunger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst*. D**52**, 57–64.

Pearlman, D. A. & Kim, S.-H. (1990). *Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data. J. Mol. Biol*. **211**, 171–187.

Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1986). Editors. *Numerical Recipes*, pp. 498–546. Cambridge University Press.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst*. A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst*. A**46**, 900–912.

Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol*. **278**, 110–128.

Rice, L. M. & Brunger, A. T. (1994). *Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. Proteins Struct. Funct. Genet*. **19**, 277–290.

Rice, L. M., Shamoo, Y. & Brunger, A. T. (1998). *Phase improvement by multi-start simulated annealing refinement and structure-factor averaging. J. Appl. Cryst*. **31**, 798–805.

Rossmann, M. G. & Blow, D. M. (1962). *The detection of sub-units within the crystallographic asymmetric unit. Acta Cryst*. **15**, 24–51.

Silva, A. M. & Rossmann, M. G. (1985). *The refinement of southern bean mosaic virus in reciprocal space. Acta Cryst*. B**41**, 147–157.

Sim, G. A. (1959). *The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetrical structures. Acta Cryst*. **12**, 813–815.

Srinivasan, R. (1966). *Weighting functions for use in the early stages of structure analysis when a part of the structure is known. Acta Cryst*. **20**, 143–144.

Suguna, K., Bott, R. R., Padlan, E. A., Subramanian, E., Sheriff, S., Cohen, G. H. & Davies, D. R. (1987). *Structure and refinement at 1.8 Å resolution of the aspartic proteinase from Rhizopus chinensis. J. Mol. Biol*. **196**, 877–900.

Verlet, L. (1967). *Computer experiments on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules. Phys. Rev*. **159**, 98–105.

International Tables for Crystallography (2012). Vol. F, Chapter 18.3, pp. 474–484.

Chapter 18.3. Structure quality and target parameters

R. A. Engh and R. Huber

18.3.1. Purpose of restraints

> Believe statistics! Shun bias! – these, we see, are two materially different laws. (Adapted from William James.) A wise man, therefore, proportions his restraints to the evidence. (Adapted from David Hume.)

If we could adequately measure the complex diffuse scattering function of an X-ray beam from a single macromolecular structure to arbitrary resolution, we could, given an accurate model for X-ray scattering, calculate a measurement-time-averaged electron-density distribution for that structure independent of most model assumptions. Alternatively, if we knew the potential energy for the system accurately as a function of all relevant parameters, classical and non-classical, and had the computational power to analyse the function appropriately, we could model the same distribution. If both experiment and theory were accessible at these extremes, the available information would be redundant and could be used together in arbitrary ways. Of course, both experiment and theory are limited far below these extremes, and the researcher must consider what information is available in order to address the pertinent protein-structure question in the most appropriate way.

The most general theoretical assumptions used to interpret experimentally determined X-ray reflections typically include the spherical symmetry of scattering functions centred at nuclei, harmonicity (and isotropy) of deviations about average positions, infinity of a crystal lattice of identical unit cells and so on. While the majority of structures determined to date share a standard set of assumptions, significant alternatives have been proposed, such as, for example, the modelling of motion with normal-mode analyses instead of harmonic temperature factors (Kidera *et al.*, 1994). Any significant change in the set of assumptions used in structure determination must be expected to leave its mark on the statistical properties of the structure.

Besides the necessity of general model assumptions for structure determination, there is the simple necessity to supplement the measured reflection data with additional sources of information to improve the ratio of experimental observables to model parameters. For example, a crystal of a protein with some 3000 atoms that diffracts to a moderate resolution of 2.5 Å might produce some 14 000 measurable unique reflections. In this case, fitting a model that includes three Cartesian coordinates and a temperature factor for each atom (12 000 parameters) is effectively underdetermined when considering experimental error. The solution is either to reduce the number of parameters by introducing constraints to the model, or to increase the number of effective observations by using additional restraints in the structure determination.

18.3.1.1. *Utility of restraints: protein/special geometries*

One set of restraints arises from the expectation that the macromolecular structure should be chemically reasonable, that is, it should reflect the geometries required by chemical physics. This restricts especially values of bond lengths, bond angles, planarity and improper dihedral angles. Additional restriction comes from the expectation that these values should most closely conform to structures determined by the same experimental method and corresponding model assumptions. There are now two sources of structural information of sufficient quality for use as restraints in structure refinement: chemical fragments from the Cambridge Structural Database (CSD) and the very high resolution protein structures that can be refined without recourse to restraining parameters (Longhi *et al.*, 1998; Dauter *et al.*, 1997). Restraints derived from the CSD have come into widespread use for protein (Engh & Huber, 1991; Brünger, 1993; Priestle, 1994) and nucleic acid (Parkinson *et al.*, 1996) refinement; since their derivation, the database has grown from some 80 000 structures to around 200 000. The number of very high resolution protein structures solved has also increased to include dozens of structures (or thousands of chemical fragments). Chemical-fragment geometries can be studied using either or both sets of structures.

The optimal choice of restraints to use depends on the character and expected use of the refined structure. For example, the quality of comparative studies (Kleywegt & Jones, 1998) is enhanced by standardized refinement conditions, removing at least one source of systematic variation (*e.g.* Laskowski, Moss & Thornton, 1993), although effects from individual crystals and data collection and processing remain. On the other hand, a structure with an unusual chemical environment might require a tailored parameterization scheme to study the effects of this environment. Such cases might occur, for example, at enzyme active sites, where bound ligands may be distorted toward intermediate geometries (Bürgi & Dubler-Steudle, 1988).

18.3.1.2. *Risk of restraints: bias, lack of cross validation*

As an alternative to implementation as restraints, statistical information from known structures provides an independent check of structural integrity. However, the more information which is used for restraints, the less is available for such cross validation. For example, the use of a force field in protein refinement that strictly enforces a physical distribution of the protein backbone φ–ψ angles of the Ramachandran plot would transfer error-induced strain into other degrees of freedom while eliminating a Ramachandran analysis as a tool for judging protein quality (Hooft *et al.*, 1997). There is an additional risk associated with the intentional introduction of bias into protein refinement: Restraining a model to conform to known structures is in error if the structure is unique in some way and therefore should *not* conform to these structures, or if there is erroneous bias in the set of structures from which the restraints were derived. Furthermore, the bias may be exaggerated, especially if the biased feature is so probable that deviations are considered suspect. This can be seen, for example, in the increasing frequency of *cis*-prolines as protein-structure resolution increases (Stewart *et al.*, 1990): when there is uncertainty in lower-resolution structures, the more probable *trans*-prolines are preferentially chosen in model building, exaggerating their frequency.

Copyright © 2012 International Union of Crystallography

18.3.2. Formulation of refinement restraints

A priori information regarding protein structure can be used in two fundamental ways: as constraints by fixing parameters to target values, or as restraints by allowing limited deviation from target values. The two methods differ fundamentally when counting the total number of degrees of freedom (important, for example, when thermodynamic quantities of simulations are considered), but both improve the observation-to-parameter ratio (constraints reduce parameters, restraints increase observations). In this chapter, we focus on the use of restraints.

A common form of restraint is an energy function parameterized to represent a conformational energy of a protein, driving the refinement toward a low-energy conformation but allowing reasonable deviations from it. Strictly, however, protein-structure refinement involves fitting model parameters to data measured from the ensemble of structures in the crystal; the model parameters should reflect ensemble properties rather than conformational energies of a single protein. This distinction may be subtle but it is statistically significant, particularly since the models themselves are simplified. For example, if the protein structure to be refined is modelled as a static structure with harmonic vibrations, then systematic errors could be expected if bond vibrations include significant anharmonicity. The systematically shortened C—H bonds of high-resolution X-ray crystal structures relative to those from neutron diffraction studies also show the systematic error involving the strictly erroneous approximation that scattering functions are spherically symmetric and centred at nuclei. These examples illustrate parameter problems that can in principle be minimized by correcting the refinement model, such as with the application of force-field restraints to an ensemble of structures, the addition of a statistically derived shift to proton positions, anisotropic scattering functions *etc.*

The 'stiffness' of the restraints should reflect the degree of confidence in the restraints. Here it is possible to distinguish different kinds of confidence, depending on the type of restraint used. The use of a conformational energy function as a refinement target function reflects the expectation that the magnitude of true deviations from the energy minima are related to the slopes of the physical energy potential surface; 'confidence' for bond lengths can thus be estimated from their vibration constants. An alternative approach is to derive target values from a database of independently determined structures. Distributions of parameter values derived from such databases have, in principle, arbitrary functional forms; a series expansion delivers the common descriptors of mean, variance, skew and so forth. The relatedness of the database to the system to be restrained determines how closely the distribution in the database should reflect the distribution in the restrained system. For proteins, it has seemed reasonable to expect that bond and angle parameters should share similar mean and sample variance distributions.

18.3.2.1. Choice of properties for restraint

Standard refinement procedures usually include the use of harmonic restraints of bond lengths, bond angles, planarity and 'improper' dihedrals, which, along with nuclear repulsions, comprise the 'hardest' restraints. Other geometric properties, such as dihedral angles, electrostatic interactions, Ramachandran energies *etc.*, can contribute further information to the refinement; these softer restraints can either distort the structure if weighted too strongly, or not contribute significantly to the refinement if weighted too weakly. Further, the softer restraints may be more useful as statistical parameters for judging the quality of a structure than as refinement parameters themselves. Particular care should be taken when refining with significant electrostatic energies: the forces have significant effects over long ranges, are usually based on simplified polarizability models and usually arise from assumptions of standard intrinsic pK_a values.

18.3.2.2. Simple derivation of force constants from parameter distributions

For a given distribution of parameter values, we take the average as the target value. If the choice of fragment geometries gives a distribution that corresponds to the varieties in the refinement problems, restraints should be applied to allow a similar range of freedom. In general, Gaussian distributions are assumed (with some obvious checks to avoid bimodal distributions) and are not corrected for non-orthogonality when parameters are introduced. If the refinement program uses a least-squares minimization method, the statistical mean and variance values can be used as tabulated. If an energy-function target is used, the variance values may be converted to force constants k defined by $E = k(\mathrm{d}x)^2$ by equating the Boltzmann probability, $\exp(-\mathrm{d}E/bT) = \exp[-k(\mathrm{d}x)^2/bT]$, with the Gaussian distribution function, $\exp[-(\mathrm{d}x)^2/(2\sigma^2)]$, such that $k = bT/2\sigma^2$, where b is the Boltzmann constant. (Note that this force constant k is one-half the magnitude of the force constant k' defined by Newton's law, $F = \mathrm{d}E/\mathrm{d}x = -k'x$.) This treatment resembles the generation of a mean field potential from the structure (*e.g.* Sippl, 1995), but is simplified for bond and angle parameters in several respects. Firstly, a single temperature is assumed; temperature dependence of equilibrium constants would require consideration of the temperatures of individual crystal structure determinations. Secondly, a normal-mode analysis would be required to eliminate redundancies arising from coupling between the parameters. Finally, the values of the variances are in many cases uncertain, due to poor statistical representation, non-Gaussian effects, or other causes (see below). Note that this simplified treatment of parameter uncertainty can be implemented more rigorously in maximum-entropy refinement methods (Pannu & Read, 1996; Bricogne, 1993).

18.3.2.2.1. Clustering

With ample computational power available for most refinement applications, the clustering of parameters into atom types might seem unnecessary. However, clustering is inevitable and occurs at least in the choice of fragments for deriving statistics. Values derived for individual residues, for example, effectively cluster three-dimensional structures together with a concomitant loss of information. For at least some residues with limited representation in the CSD, peptide geometries are best analysed as a general class, requiring clustering fragments into side-chain and main-chain classifications. Some side chains, such as arginine, also require smaller fragment definitions due to relatively small numbers in the CSD. The statistics listed here can be further clustered into fewer atom or bond-type classes. Inappropriate clustering, that is, the simultaneous analysis of fragments that are better represented by two or more fragment types with correspondingly various average values and sample variances, will exaggerate non-Gaussian distribution characteristics. In extreme cases, skew, kurtosis and especially multimodal distributions then provide evidence for a requirement to subdivide the fragment classes (see for example *cis*/*trans*-proline below). The effective clustering of the data presented here into largely conformation-

Table 18.3.2.1
Bond lengths of standard amino-acid side chains

EH denotes the values of Engh & Huber (1991), which were clustered according to atom type. The EH99 values are taken from recent Cambridge Structural Database releases with clustering of parameters only in the choice of fragments, based on amino acids. Parameters marked with an asterisk involving CA—CB bonds were taken from peptide fragment geometries. Two asterisks mark long-chain aliphatic parameters taken from arginine statistics. The number of fragments and the number of structures containing these fragments are noted after the amino-acid name. The fragments used for generating the statistics are described after the amino-acid name: incomplete valences indicate unspecified substituents with, however, specified orbital hybridization.

Alanine, 163/268, $CO-NH-CH(CH_3)-CO-NH$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.521	0.033	1.520	0.021

Arginine, 71/98, $CH-(CH_2)_3-NH-C(NH_2)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521	0.027
CG—CD	1.520	0.030	1.515	0.025
CD—NE	1.460	0.018	1.460	0.017
NE—CZ	1.329	0.014	1.326	0.013
CZ—NH(1,2)	1.326	0.018	1.326	0.013

Asparagine, 145/247, $-C-CH_2-CO-NH_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.527	0.026
CB—CG	1.516	0.025	1.506	0.023
CG—OD1	1.231	0.020	1.235	0.022
CG—ND2	1.328	0.021	1.324	0.025

Aspartate, 265/404, $C-CH_2-CO_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.516	0.025	1.513	0.021
CG—OD(1,2)	1.249	0.019	1.249	0.023

Cysteine, 10/17, $N-CH(CO)-CH_2-SH$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.526	0.013
CB—SG	1.808	0.033	1.812	0.016

Disulfides, 53/68, $C-CH_2-S-S-CH_2-C$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—SG	1.808	0.033	1.818	0.017
SG—SG	2.030	0.008	2.033	0.016

Glutamate, 74/88, $C-CH_2-CH_2-CO_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.517	0.019
CG—CD	1.516	0.025	1.515	0.015
CD—OE(1,2)	1.249	0.019	1.252	0.011

Glutamine, 145/247, $-C-CH_2-CO-NH_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521**	0.027**
CG—CD	1.516	0.025	1.506	0.023
CD—OE1	1.231	0.020	1.235	0.022
CD—NE2	1.328	0.021	1.324	0.025

Glycine: see peptide parameters, Table 18.3.2.3

Histidine (HISE), 35/37, $C-CH_2-$imidazole; NE protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.496	0.018
CG—ND1	1.371	0.017	1.383	0.022
CG—CD2	1.356	0.011	1.353	0.014
ND1—CE1	1.319	0.013	1.323	0.015
CD2—NE2	1.374	0.021	1.375	0.022
CE1—NE2	1.345	0.020	1.333	0.019

Histidine (HISD), 10/12, $C-CH_2-$imidazole; ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.016
CG—ND1	1.378	0.011	1.369	0.015
CG—CD2	1.356	0.011	1.353	0.017
ND1—CE1	1.345	0.020	1.343	0.025
CD2—NE2	1.382	0.030	1.415	0.021
CE1—NE2	1.319	0.013	1.322	0.023

Histidine (HISH), 50/54, $C-CH_2-$imidazole; NE, ND protonated

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.497	0.014	1.492	0.010
CG—ND1	1.378	0.011	1.380	0.010
CG—CD2	1.354	0.011	1.354	0.009
ND1—CE1	1.321	0.010	1.326	0.010
CD2—NE2	1.374	0.011	1.373	0.011
CE1—NE2	1.321	0.010	1.317	0.011

Isoleucine, 54/80, $NH-CH(CO)-CH(CH_3)-CH_2-CH_3$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.544	0.023
CB—CG1	1.530	0.020	1.536	0.028
CB—CG2	1.521	0.033	1.524	0.031
CG1—CD1	1.513	0.039	1.500	0.069*

Leucine, 178/288, $NH-CH(CO)-CH_2-CH(CH_3)_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.533	0.023
CB—CG	1.530	0.020	1.521	0.029
CG—CD(1,2)	1.521	0.033	1.514	0.037

Lysine, 232/380, $-(CH_2)_3-NH_3$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.521**	0.027**
CG—CD	1.520	0.030	1.520	0.034
CD—CE	1.520	0.030	1.508	0.025
CE—NZ	1.489	0.030	1.486	0.025

Methionine, 37/49, $C-(CH_2)_2-S-CH_3$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.520	0.030	1.509	0.032
CG—SD	1.803	0.034	1.807	0.026
SD—CE	1.791	0.059	1.774	0.056*

Table 18.3.2.1 (continued)

Phenylalanine, 1076/1616, C—$\mathrm{CH_2}$—phenyl

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.502	0.023	1.509	0.017
CG—CD(1,2)	1.384	0.021	1.383	0.015
CD(1,2)—CE(1,2)	1.382	0.030	1.388	0.020
CE(1,2)—CZ	1.382	0.030	1.369	0.019

Proline, 262/255, *trans*, C—CO—pyrrolidine—CO—N

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.531	0.020
CB—CG	1.492	0.050	1.495	0.050
CG—CD	1.503	0.034	1.502	0.033
CD—N	1.473	0.014	1.474	0.014

Proline, 262/158, *cis*, C—CO—pyrrolidine—CO—N

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.533	0.018
CB—CG	1.492	0.050	1.506	0.039
CG—CD	1.503	0.034	1.512	0.027
CD—N	1.473	0.014	1.474	0.014

Serine, 33/39, NH—CH(CO)—$\mathrm{CH_2}$—OH

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.525	0.015
CB—OG	1.417	0.020	1.418	0.013

Threonine, 20/25, NH—CH(CO)—CH(OH)—$\mathrm{CH_3}$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.529	0.026
CB—OG1	1.433	0.016	1.428	0.020
CB—CG2	1.521	0.033	1.519	0.033

Tryptophan, 123/135, $\mathrm{CH_2}$—indole

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.498	0.031	1.498	0.018
CG—CD1	1.365	0.025	1.363	0.014
CG—CD2	1.433	0.018	1.432	0.017
CD1—NE1	1.374	0.021	1.375	0.017
NE1—CE2	1.370	0.011	1.371	0.013
CD2—CE2	1.409	0.017	1.409	0.012
CD2—CE3	1.398	0.016	1.399	0.015
CE2—CZ2	1.394	0.021	1.393	0.017
CE3—CZ3	1.382	0.030	1.380	0.017
CZ2—CH2	1.368	0.019	1.369	0.019
CZ3—CH2	1.400	0.025	1.396	0.016

Tyrosine, 124/161, *para*-(—C—$\mathrm{CH_2}$)—phenol

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.530	0.020	1.535*	0.022*
CB—CG	1.512	0.022	1.512	0.015
CG—CD(1,2)	1.389	0.021	1.387	0.013
CD(1,2)—CE(1,2)	1.382	0.030	1.389	0.015
CE(1,2)—CZ	1.378	0.024	1.381	0.013
CZ—OH	1.376	0.021	1.374	0.017

Valine, 198/313, N—CH(CO)—CH—$(\mathrm{CH_3})_2$

Bond	EH (Å)	σ EH (Å)	EH99 (Å)	σ EH99 (Å)
CA—CB	1.540	0.027	1.543	0.021
CB—CG(1,2)	1.521	0.033	1.524	0.021

independent fragment definitions could be replaced by more specific information, especially as the database grows.

18.3.2.2.2. *Treatment of outliers*

A truly Gaussian distribution should include outliers at high σ values (about 0.01% for 4σ). We should expect, however, that the width of the distribution is affected not only by inherent variation in the variables to be parameterized, but also by variability in the experimental conditions (*e.g.* resolution) and by erroneous structures. This weakens a strategy of automatic rejection of outliers beyond a specific cutoff value. The possibility of visualizing the distributions with CSD software allows refinement of this rejection strategy, with, however, the introduction of considerable subjectivity in the criteria. For this work, a 4σ cutoff was generally considered a flag for erroneous outliers. However, broad and flat tails in the distribution were relatively frequent and often asymmetric. These deviations from Gaussian behaviour 'artificially' increased σ values. In these cases, the 4σ cutoff rule was not applied automatically, but was applied after examination and rejection of conspicuous outliers. From an algorithmic viewpoint, this was the additional use of skew and kurtosis (third and fourth moments of the distribution) for rejection criteria. In most cases, uncertainty in rejection criteria affected the average values little, but could significantly alter standard deviations.

18.3.2.3. *Bonds and angles*

18.3.2.3.1. *Peptide parameters: proline, glycine, alanine and CB substitution*

Fragments representing five-atom lengths of the backbone currently provide adequate statistics for peptide compositions of varieties including glycine, proline and side chains branched at CB. Peptide cyclicity was generally allowed on the assumption that this does not introduce distortions greater than typical protein secondary-structure interactions. The results are presented in Table 18.3.2.3. With one exception, none of the values deviates from those of 1991 by more than one sample standard deviation. However, the very large σ values for the proline C—N—CA and C—N—CD angles (Table 18.3.2.1) are conspicuous. Using high-resolution protein structures, Lamzin *et al.* (1995) identified geometries of proline that were inconsistent with high-resolution protein structures and also noted inconsistencies in C—CA—CB angle parameters (see also the sections on individual amino acids below). In the case of proline, a bimodal distribution of these parameters could be resolved with the discrimination between *cis* and *trans* forms (Fig. 18.3.2.1). A scatter plot of the angles against ω torsion angle resolves the averages (and σ's) of 122.6 (50) and 125.4 (44)° for C—N—CA and C—N—CD, respectively, into *cis*- and *trans*-dependent values with much smaller sample deviations (see Table 18.3.2.2). The large σ value for CB—CG remains, however, particularly for *trans*-proline. Its origin is unknown, but proline pucker may play a role.

Glycine, with its unique $\mathrm{CH_2}$ as CA, required new atom-type definitions for Engh & Huber (EH) (1991) parameterization to account for parameter-average differences of about one-half of a sample standard deviation. These also included C—N—CA, for which the average angles were 120.6° for glycine and 121.7° for the rest. The new statistics with 83 C—CO—NH—$\mathrm{CH_2}$—C fragments estimate a larger value of 122.3° for the glycine C—N—CA angle.

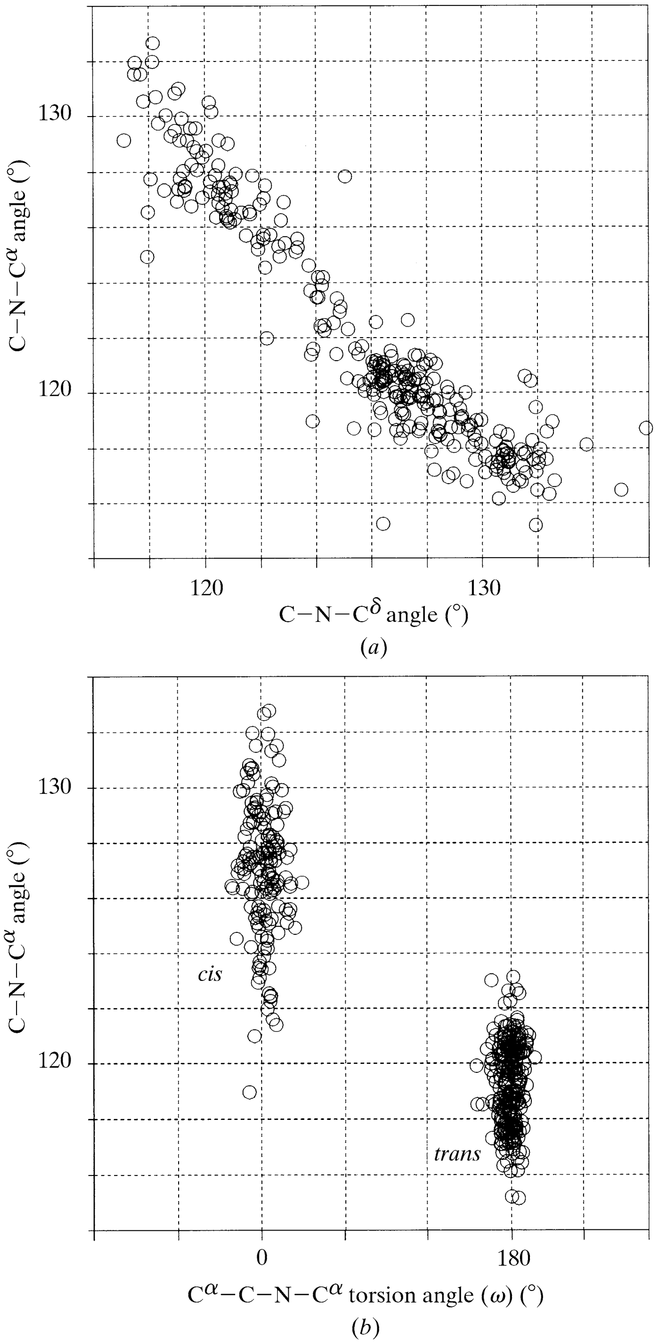


Figure 18.3.2.1
Torsion dependence of proline angle geometry. A one-dimensional frequency plot of either C—N—C$^{\alpha}$ or C—N—C$^{\delta}$ angles shows a broad and bimodal distribution. (*a*) A scatter plot of the two angles shows a very strong anticorrelation and suggests two minima. (*b*) Plotting either angle against the ω torsion angle resolves the broad distribution into two separate peaks.

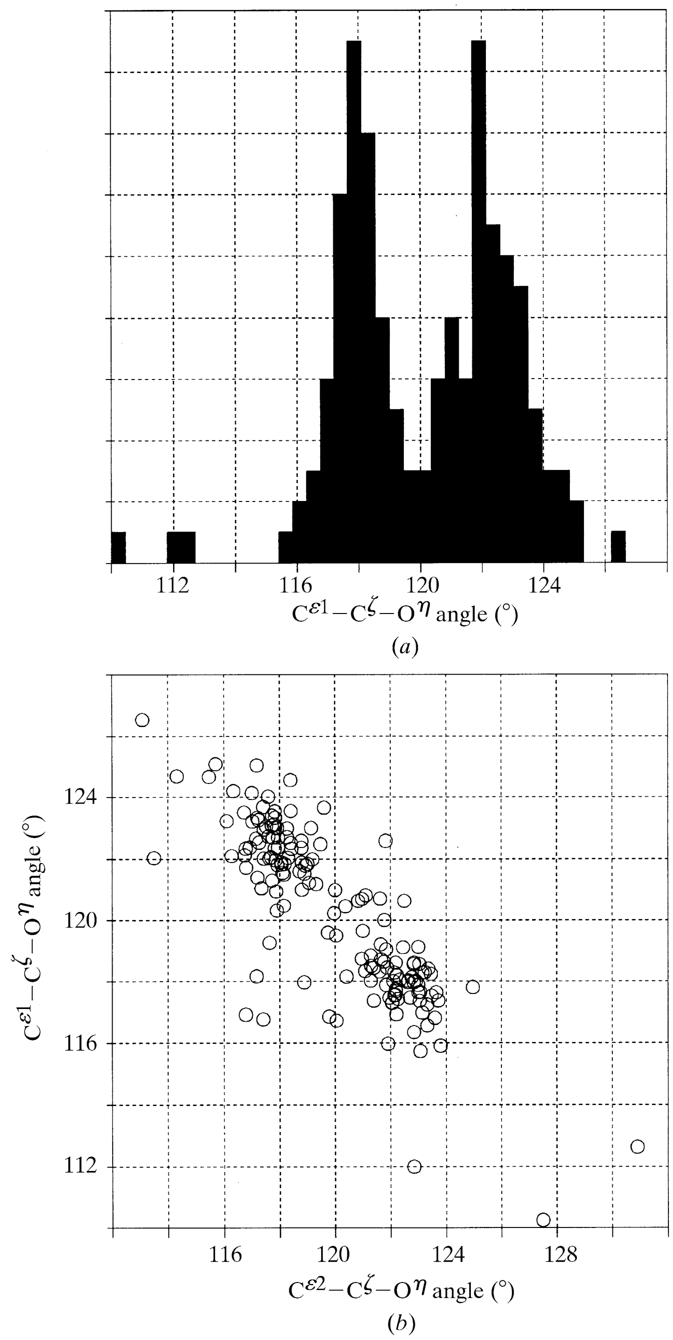


Figure 18.3.2.2
Bimodal distributions for tyrosine. (*a*) The C$^{\varepsilon}$—C$^{\zeta}$—O$^{\eta}$ angle distributions involving the tyrosine alcohol have maxima at 120 (2)°. (*b*) A scatter plot of the C$^{\varepsilon 2}$—C$^{\zeta}$—O$^{\eta}$ angle against the C$^{\varepsilon 1}$—C$^{\zeta}$—O$^{\eta}$ angle confirms that the C$^{\zeta}$—O$^{\eta}$ bond projects asymmetrically away from the aromatic ring.

'Extended atom'-type parameterizations, which cluster carbon atoms according to the number of bound hydrogen atoms, naturally separate parameters involving CB into values representing alanine and branched and unbranched side chains. Separate analyses of the bonds and angles for fragments depending on the number of hydrogen atoms at CB (1, 2 or 3) revealed significant variation for the C—CA—CB and N—CA—CB angles. The fragments chosen for peptide parameterization did not cover all possibilities for the peptide chain. In particular, effects of charges at the termini were not analysed. Also, specific residue sequences likely to have statistical effects, such as Pro-Pro (Bansal & Ananthanarayanan, 1988), were not analysed here. With 50–60 relevant fragments from the predominantly α-helical ROP protein, Vlassi *et al.* (1998) were able to compile statistics for main-chain bonds and angles and compare them with protein refinement parameters. Differences from EH were particularly significant for CO and CA—C bonds (1.237 and 1.508 Å, respectively) and for the O—C—N angle (121.35°). Excepting the proline O—C—N angle, for which the new CSD statistics predict an average value lowered to 121.1°, these values remained relatively unchanged. A likely source of the difference might be the predominantly helical structure of the ROP protein; the helical hydrogen bonding directly involves the C—O group in a systematic way.

Table 18.3.2.2
Bond angles of standard amino-acid side chains

For details see Table 18.3.2.1.

Alanine, 163/268, CO—NH—CH(CH$_3$)—CO—NH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.4	1.5	110.1	1.4
CB—CA—C	110.5	1.5	110.1	1.5

Arginine, 71/98, CH—(CH$_2$)$_3$—NH—C(NH$_2$)$_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4	2.2
CB—CG—CD	111.3	2.3	111.6	2.6
CG—CD—NE	112.0	2.2	111.8	2.1
CD—NE—CZ	124.2	1.5	123.6	1.4
NE—CZ—NH(1,2)	120.0	1.9	120.3	0.5
NH1—CZ—NH2	119.7	1.8	119.4	1.1

Asparagine, 145/247, —C—CH$_2$—CO—NH$_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	112.6	1.0	113.4**	2.2**
CB—CG—ND2	116.4	1.5	116.7	2.4
CB—CG—OD1	120.8	2.0	121.6	2.0
ND2—CG—OD1	122.6	1.0	121.9	2.3

Aspartate, 265/404, C—CO$_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	112.6	1.0	113.4*	2.2**
CB—CG—OD(1,2)	118.4	2.3	118.3	0.9
OD1—CG—OD2	122.9	2.4	123.3	1.9

Cysteine, 10/17, N—CH(CO)—CH$_2$—SH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.8	1.5
CB—CA—C	110.1	1.9	111.5	1.2
CA—CB—SG	114.4	2.3	114.2	1.1

Disulfides, 53/68, C—CH$_2$—S—S—CH$_2$—C

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—SG	114.4	2.3	114.0	1.8
CB—SG—SG	103.8	1.8	104.3	2.3

Glutamate, 74/88, C—CH$_2$—CH$_2$—CO$_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4**	2.2**
CB—CG—CD	112.6	1.7	114.2	2.7
CG—CD—OE(1,2)	118.4	2.3	118.3	2.0
OE1—CD—OE2	122.9	2.4	123.3	1.2

Glutamine, 145/247, —C—CH$_2$—CO—NH$_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4**	2.2**
CB—CG—CD	112.6	1.7	111.6**	2.6**
CG—CD—OE1	120.8	2.0	121.6	2.0
CG—CD—NE2	116.4	1.5	116.7	2.4
OE1—CD—NE2	122.6	1.0	121.9	2.3

Glycine: see Table 18.3.2.3

Histidine (HISE), 35/37, C—CH$_2$—imidazole; NE protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.7
CB—CG—ND1	121.6	1.5	121.4	1.3
CB—CG—CD2	129.1	1.3	129.7	1.6
CG—ND1—CE1	105.6	1.0	105.7	1.3
ND1—CE1—NE2	111.7	1.3	111.5	1.3
CE1—NE2—CD2	106.9	1.3	107.1	1.1
NE2—CD2—CG	106.5	1.0	106.7	1.2
CD2—CG—ND1	109.2	0.7	108.8	1.4

Histidine (HISD), 10/12, C—CH$_2$— imidazole; ND protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.7
CB—CG—ND1	122.7	1.5	123.2	2.5
CB—CG—CD2	129.1	1.3	130.8	3.1
CG—ND1—CE1	109.0	1.7	108.2	1.4
ND1—CE1—NE2	111.7	1.3	109.9	2.2
CE1—NE2—CD2	107.0	3.0	106.6	2.5
NE2—CD2—CG	109.5	2.3	109.2	1.9
CD2—CG—ND1	105.2	1.0	106.0	1.4

Histidine (HISH), 50/54, C—CH$_2$—imidazole; NE, ND protonated

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.6	1.6
CB—CG—ND1	122.7	1.5	122.5	1.3
CB—CG—CD2	131.2	1.3	131.4	1.2
CG—ND1—CE1	109.3	1.7	109.0	1.0
ND1—CE1—NE2	108.4	1.0	108.5	1.1
CE1—NE2—CD2	109.0	1.0	109.0	0.7
NE2—CD2—CG	107.2	1.0	107.3	0.7
CD2—CG—ND1	106.1	1.0	106.1	0.8

Isoleucine, 54/80, NH—CH(CO)—CH(CH$_3$)—CH$_2$—CH$_3$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	110.8	2.3
CB—CA—C	109.1	2.2	111.6	2.0
CA—CB—CG1	110.4	1.7	111.0	1.9
CB—CG1—CD1	113.8	2.1	113.9	2.8
CA—CB—CG2	110.5	1.7	110.9	2.0
CG1—CB—CG2	110.7	3.0	111.4	2.2

Table 18.3.2.2 (continued)

Leucine, 178/288, NH—CH(CO)—CH_2—$CH(CH_3)_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.4	2.0
CB—CA—C	110.1	1.9	110.2	1.9
CA—CB—CG	116.3	3.5	115.3	2.3
CB—CG—CD(1,2)	110.7	3.0	111.0	1.7
CD1—CG—CD2	110.8	2.2	110.5	3.0

Lysine, 232/380, —$(CH_2)_3$—NH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.4*	2.2**
CB—CG—CD	111.3	2.3	111.6**	2.6**
CG—CD—CE	111.3	2.3	111.9	3.0
CD—CE—NZ	111.9	3.2	111.7	2.3

Methionine, 37/49, C—$(CH_2)_2$—S—CH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	114.1	2.0	113.3	1.7
CB—CG—SD	112.7	3.0	112.4	3.0
CG—SD—CE	100.9	2.2	100.2	1.6

Phenylalanine, 1076/1616, C—CH_2—phenyl

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.8	1.0	113.9	2.4
CB—CG—CD(1,2)	120.7	1.7	120.8	0.7
CD(1,2)—CG—CD(2,1)	118.6	1.5	118.3	1.3
CG—CD(1,2)—CE(1,2)	120.7	1.7	120.8	1.1
CD(1,2)—CE(1,2)—CZ	120.0	1.8	120.1	1.2
CE(1,2)—CZ—CE(2,1)	120.0	1.8	120.0	1.8

Proline, 262/255, *trans*, C—CO—pyrrolidine—CO—N

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	103.0	1.1	103.3	1.2
CB—CA—C	110.1	1.9	111.7	2.1
CA—CB—CG	104.5	1.9	104.8	1.9
CB—CG—CD	106.1	3.2	106.5	3.9
CG—CD—N	103.2	1.5	103.2	1.5
CA—N—CD	112.0	1.4	111.7	1.4
C—N—CA	122.6	5.0	119.3	1.5
C—N—CD	125.0	4.1	128.4	2.1

Proline, 262/158, *cis*, C—CO—pyrrolidine—CO—N

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	103.0	1.1	102.6	1.1
CB—CA—C	110.1	1.9	112.0	2.5
CA—CB—CG	104.5	1.9	104.0	1.9
CB—CG—CD	106.1	3.2	105.4	2.3
CG—CD—N	103.2	1.5	103.8	1.2
CA—N—CD	112.0	1.4	111.5	1.4
C—N—CA	122.6	5.0	127.0	2.4
C—N—CD	125.0	4.1	120.6	2.2

Serine, 33/39, NH—CH(CO)—CH_2—OH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.5	1.5
CB—CA—C	110.1	1.9	110.1	1.9
CA—CB—OG	111.1	2.0	111.2	2.7

Threonine, 20/25, NH—CH(CO)—CH(OH)—CH_3

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	110.3	1.9
CB—CA—C	109.1	2.2	111.6	2.7
CA—CB—OG1	109.6	1.5	109.0	2.1
CA—CB—CG2	110.5	1.7	112.4	1.4
OG1—CB—CG2	109.3	2.0	110.0	2.3

Tryptophan, 123/135, CH_2—indole

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.6	1.9	113.7†	1.9†
CB—CG—CD1	126.9	1.5	127.0	1.3
CB—CG—CD2	126.8	1.4	126.6	1.3
CD1—CG—CD2	106.3	1.6	106.3	0.8
CG—CD1—NE1	110.2	1.3	110.1	1.0
CD1—NE1—CE2	108.9	1.8	109.0	0.9
NE1—CE2—CD2	107.4	1.3	107.3	1.0
CE2—CD2—CG	107.2	1.2	107.3	0.8
CG—CD2—CE3	133.9	1.0	133.9	0.9
NE1—CE2—CZ2	130.1	1.5	130.4	1.1
CE3—CD2—CE2	118.8	1.0	118.7	1.2
CD2—CE2—CZ2	122.4	1.0	122.3	1.2
CE2—CZ2—CH2	117.5	1.3	117.4	1.0
CZ2—CH2—CZ3	121.5	1.3	121.6	1.2
CH2—CZ3—CE3	121.1	1.3	121.2	1.1
CZ3—CE3—CD2	118.6	1.3	118.8	1.3

Tyrosine 124/161, *para*—C—CH_2—phenyl—OH

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	110.5	1.7	110.6*	1.8*
CB—CA—C	110.1	1.9	110.4*	2.0*
CA—CB—CG	113.9	1.8	113.4	1.9
CB—CG—CD(1,2)	120.8	1.5	121.0	0.6
CD(1,2)—CG—CD(2,1)	118.1	1.5	117.9	1.1
CG—CD(1,2)—CE(1,2)	121.2	1.5	121.3	0.8
CD(1,2)—CE(1,2)—CZ	119.6	1.8	119.8	0.9
CE(1,2)—CZ—CE(2,1)	120.3	2.0	119.8	1.6
CE(1,2)—CZ—OH	119.9	3.0	120.1	2.7†

Valine, 198/313, N—CH(CO)—CH—$(CH_3)_2$

Angle	EH (°)	σ EH (°)	EH99 (°)	σ EH99 (°)
N—CA—CB	111.5	1.7	111.5	2.2
CB—CA—C	109.1	2.2	111.4	1.9
CA—CB—CG(1,2)	110.5	1.7	110.9	1.5
CG1—CB—CG2	110.8	2.2	110.9	1.6

† Alternate fragment definition including CA. † Bimodal distribution (see text).

18.3.2.3.2. Aromatic residues: tryptophan, phenylalanine, tyrosine, histidine

With the exception of generally lower σ values, tryptophan parameters remain essentially unchanged. Phenylalanine, also with generally lower σ values, is also essentially unchanged with the assumption of Gaussian distributions. However, a scatter plot of the CB—CG—CD1 *versus* CB—CG—CD2 angles shows an inverse correlation between these two angles, corresponding to ring rotations about an axis perpendicular to the ring face. Non-Gaussian distributions were most evident for tyrosine. In addition to the phenomenon described for phenylalanine, a clearly multimodal distribution was observed for the CE(1,2)—CZ—OH angles, with maxima at 118 and 122° (Fig. 18.3.2.2). The scatter plot of CE1—CZ—OH *versus* CE2—CZ—OH demonstrates that this distribution typifies individual fragments and does not arise from differing classes of fragments. This justifies an asymmetric parameterization for these angles; symmetric parameterization would require correspondingly soft force constants. The major difference between the histidine parameters listed here compared to those of EH arise from the appearance of HISD (uncharged; unprotonated at NE2) fragments in the CSD. The EH parameterization assumed values from other fragments. The total of 12 fragments is not large, but does predict some alterations in parameters involving the ring nitrogens. The fragment selection reported here did not investigate effects of noncovalent binding. For the aromatic residues, these include hydrogen-bonding effects (especially for histidine) and π-cloud interactions. Appropriate fragments exist in the database, so such dependencies are, in principle, accessible to investigation.

18.3.2.3.3. Aliphatic residues: leucine, isoleucine, valine

Compared to EH parameterization, the only notable features of the aliphatic residues were the leucine bonds and the C—CA—CB angles of isoleucine and valine. The leucine CD—CG(1,2) bonds retained relatively large σ values, which rather increased compared to the previous values. The C—CA—CB angle values, clustered as bare carbon/tetrahedral CH extended atom/tetrahedral CH_2 extended atom in EH, are sensitive to the degree of substitution at the CB carbon (Table 18.3.2.3, see the discussion of peptide fragments above). The statistics here show that the EH (1991) parameters were too small by about 2°.

18.3.2.3.4. Neutral polar residues: serine, threonine, glutamine, asparagine

These residues share neutral polarity, but are all geometrically distinct. Like leucine, valine and isoleucine described above, threonine is branched at CB, and the parameterization for C—CA—CB should be chosen accordingly. Additionally for threonine, the CA—CB—CG2 angle, clustered with valine as CH1E—CH1E—CH3E in EH (1991), should be altered from 110.5 to 112.4° according to the statistics reported here. The tabulated glutamine and asparagine parameters are taken from identical amide-group statistics, and parameters for the aliphatic atoms of glutamine are taken from arginine. This choice of fragments arose from a desire to maximize the number of fragments for the amide group; however, the individual residues might be expected to exhibit residue-specific amide structures.

18.3.2.3.5. Acidic residues: glutamate, aspartate

The fragment definitions were chosen to select both symmetrically and asymmetrically encoded carboxylate structures; that is, the statistics include carboxylate groups with delocalized charges as well as carboxylate groups encoded with a single charged oxygen atom. This distribution presumably reflects the variations in proteins as well. For both glutamic and aspartic acids, statistical variation in the asymmetry of delocalization was evident. One measure of parameter variation as a function of varying charge delocalization is the anticorrelation of C—O bond lengths and CH_2—C—O bond angles. For example, while the standard deviation of the corresponding aspartate bond lengths individually is 0.024 Å, the standard deviation of their pairwise average is 0.012 Å. Similarly, the standard deviation of the glutamate CH_2—C—O bond angles individually is 2.1° but the standard deviation of the pairwise average is 0.6°. This coupling of parameters is an example of additional information potentially available for structure refinement, but which would require new formulations of restraints.

18.3.2.3.6. Basic residues: arginine, lysine

The 98 arginine fragments in the database did not show alterations from the EH values, except generally tighter restraints at the guanidinium group. Lysine CD—CE bond lengths are somewhat shorter in the new statistics, while the two angles derived from the fragments remained similar.

18.3.2.3.7. Sulfur-containing residues: methionine, cysteine, disulfides

One of the most conspicuous features of the EH parameters is the soft force constant for the methionine SD—CE bond length. The 49 fragments now in the CSD also show a sample deviation for the 1.774 Å average bond length of 0.056 Å, and after one 4σ outlier rejection, the tabulated value of 1.779 Å still has a large sample deviation of 0.041 Å. In practice, the use of soft restraints during refinement often leads to warnings of relatively large deviations from the target value. Inspection of the CSD structures did not reveal an artificial source of this greater variability. Cysteines and disulfides here show reduced sample σ values for generally similar average target values.

18.3.2.4. Planarity restraints

Planarity and improper dihedral restraints, being ‘hard’ restraints, are amenable to the same kind of parameterization described above. Physically realistic deviations should be allowed. A survey of several planar atoms, such as CG of aromatic residues, the inter-ring carbons (CD2, CE2) of tryptophan and CZ of tyrosine, showed standard deviations about strict planarity of 1–2°. Statistically significant deviations of average values from perfect planarity might also be expected, particularly as a function of the protein fold environment. For example, an average nonzero planarity of the ω angle of the peptide bond has

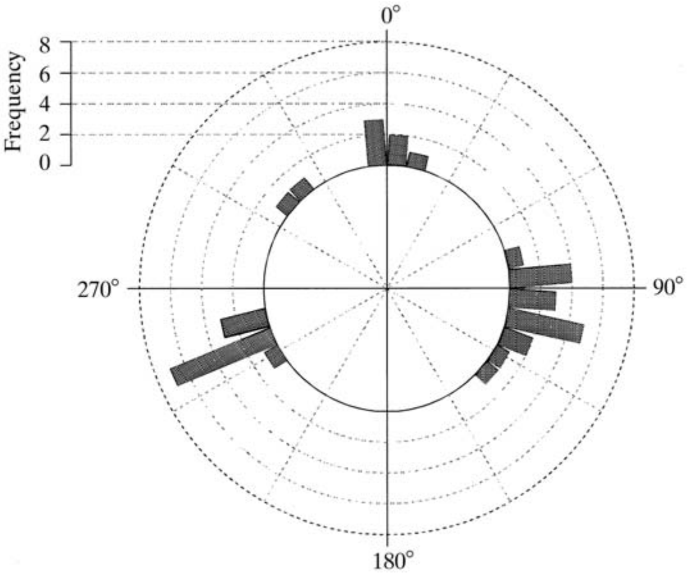


Figure 18.3.2.3

Tryptophan χ_2 dihedral angle distribution. The 36 tryptophan fragments in the CSD show several apparent minima, including an eclipsed $C^{\alpha}—C^{\beta}—C^{\gamma}—C^{\delta 1}$ dihedral conformation. This is apparent in χ_1, χ_2 tryptophan distribution plots from protein structures as well (Laskowski, MacArthur *et al.*, 1993).

Table 18.3.2.3
Bond lengths (Å) and angles (°) of peptide backbone fragments

EH denotes parameters from Engh & Huber (1991). Bold values mark important updates for angles involving proline (with *cis* and *trans* distinction) and branched CB atoms (isoleucine, valine, threonine). The number of fragments used for the statistics is given. The standard deviation of each value is given in parentheses following the value and is on the scale of the least significant digit of the value.

(*a*) Bonds

	Peptide (1165)	Proline (297)	Glycine (83)	EH	EH proline	EH glycine
N−CA	1.459 (20)	1.468 (17)	1.456 (15)	1.458 (19)	1.466 (15)	1.451 (16)
CA−C	1.525 (26)	1.524 (20)	1.514 (16)	1.525 (21)	EH	1.516 (18)
C−O	1.229 (19)	1.228 (20)	1.232 (16)	1.231 (20)	EH	EH
CA−CB (all)	1.532 (31)	1.531 (20)	–	1.530 (20)	EH	–
CA−CB (CH_3)	1.521 (33)	–	–	1.521 (33)	–	–
CA−CB (CH_2)	1.535 (22)	1.531 (20)	–	1.530 (20)	EH	–
CA−CB (CH)	1.542 (23)	–	–	1.540 (27)	–	–
C−N	1.336 (23)	1.338 (19)	1.326 (18)	1.329 (14)	1.341 (16)	EH

(*b*) Angles

	Peptide (1165)	Proline (297)	Glycine (83)	EH	EH proline	EH glycine
N−CA−C	111.0 (27)	112.1 (26)	113.1 (25)	111.2 (28)	111.8 (25)	112.5 (29)
N−CA−CB (all)	110.6 (21)	103.1 (12)	–	110.5 (17)	103.0 (11)	–
N−CA−CB (CH_3)	110.4 (15)	–	–	110.4 (15)	EH	–
N−CA−CB (CH_2)	110.6 (18)	103.1 (12)	–	110.5 (17)	103.0 (11)	–
N−CA−CB (CH)	111.1 (23)	–	–	111.5 (17)	EH	–
CA−C−N	117.2 (22)	117.1 (28)	116.2 (20)	116.2 (20)	116.9 (15)	116.4 (21)
CA−C−O	120.1 (21)	120.2 (24)	120.6 (18)	120.8 (17)	EH	120.8 (21)
O−C−N	122.7 (16)	121.1 (19)	123.2 (17)	123.0 (16)	122.0 (14)	EH
C−CA−CB	110.6 (23)	111.8 (20)	–	110.1 (19)	EH	–
C−CA−CB (CH_3)	110.5 (15)	–	–	110.5 (15)	–	–
C−CA−CB (CH_2)	110.4 (20)	111.8 (20)	–	110.1 (19)	EH	–
C−CA−CB (CH)	**111.3 (20)**	–	–	109.1 (22)	–	–
C−N−CA	121.7 (25)	122.0 (42) all **119.3 (15) *trans*** **127.0 (24) *cis***	**122.3 (21)**	121.7 (18)	122.6 (50)	120.6 (17)

been noted (Marquart *et al.*, 1983) and attributed at least in part to secondary structure; such effects may be misrepresented in statistics from small molecules.

18.3.2.5. *Torsion angles*

Since torsion angles are generally more adequately determined by protein structures of typical resolutions, there is less need to derive parameters from the CSD for refinement purposes. Further, distributions derived from small molecules may not be representative of torsion angles among proteins, since typical fragments lack 'typical' protein secondary structure. This kind of environmental dependence will affect softer parameters such as dihedral angles more than bonds or angles. However, since the torsion-angle distribution is a function not only of potential interactions of the peripheral groups but of the electronic character of the bond itself, some questions may be best examined by the study of chemical fragments. One such feature might be the $\chi_1\chi_2$ distributions for aromatic residues. Statistical distributions of side-chain orientations show the apparent stability of an eclipsed χ_2 conformation, particularly for $\chi_1 = 300°$ (g^- conformer, Fig. 18.3.2.3). If the relative dearth of secondary structure leads to atypical torsion-angle distributions in proteins, it must be conversely expected that torsion-angle statistics derived from the set of all proteins will be less applicable, for example, to helical proteins than statistics derived from mostly helical proteins. This illustrates the dilemma of selecting the ideal statistical database for protein refinement.

18.3.2.6. *Non-bonded interactions*

Like the potential parameterization of torsion-angle statistics with the CSD, parameterization of non-bonded interactions, typically into terms representing packing (an empirical mix of London dispersion forces and solvent effects), electrostatics and hydrogen bonding, is probably more strongly influenced by protein environment than are bond and angle terms. Specific chemical questions are likely to be best addressed with fragment structure databases, however, for improved parameterization or structure-quality evaluation. Possible examples include hydrogen bonds, salt bridges (see, *e.g.*, the discussion on charge delocalization of Glu and Asp above) *etc.* These are particularly relevant for structures generally atypical among proteins, such as at enzyme reactive sites.

18.3.2.7. *Effects of hydrogen atoms in parameterization*

While many CSD fragments include hydrogen-atom positions, their accuracy is necessarily the most limited. Evaluation of CSD statistics without hydrogens and the subsequent addition of parameters to refine hydrogens adds an additional artifactual coupling between parameters involving non-hydrogen atoms as well. This artifactual coupling might theoretically be of some concern, but is a second-order effect and presumably introduces effects smaller than the artifactual coupling between non-hydrogen parameters, for example.

18.3.2.8. *Special geometries: cofactors, ligands, metals etc.*

Most crystallographers will experience neither the need nor desire to derive their own parameterization for general protein structure refinement; many, however, need new parameters for ligands or other entities that are not amino-acid residues. The accuracy required for their application will determine the appropriate effort. If the purpose of the structure is to determine the general orientation of an inhibitor for molecular modelling studies with a still lower effective accuracy, it may be that no refinement (and no parameterization) is necessary at all. As the accuracy requirements increase, so does the need for good parameterization and a way of estimating when the density is incompatible with known structures. Such incompatibility may be decisive in identifying the stereochemistry of ligands selected from racemic mixtures, or the occurrence of a chemical reaction, or even falsely characterized substances. For such applications, small-molecule structural databases will remain the only choice for parameter derivation, which can be done exactly as for amino-acid fragments.

18.3.2.9. Addition of tailored information sources

When specific structural effects are observed which are not otherwise parameterized, new parameters might be desirable to encode this information. In the case of new statistical minima or variances of parameters already encoded by the refinement program, it is for most programs a simple matter to introduce new atom types, residues, or fragment names to the topology and parameter libraries. Examples include new parameterization for charged states or for *cis*- and *trans*-proline. If the reason for the new statistics is structural, the structure must be appropriately monitored during refinement to ensure that the conditions continue to hold, particularly in the case of simulated-annealing refinement steps.

18.3.3. Strategy of application during building/refinement

Refinement parameters necessarily and intentionally introduce bias into the refinement which may not disappear with later alterations of parameters. The importance of this fact is reflected by the observation that the parameters of refined structures can be recognized by statistical studies of the structures (Laskowski, Moss & Thornton, 1993). It is therefore important that the parameters initially reflect what can be confidently predicted about the structure. If unknown geometries may be expected, at metal or catalytic sites, for example, or if isomerization states need to be recognized from the refined structure, all relevant parameters must be initially eliminated from the refinement. Depending on the resolution of the structure and the detail required, the unbiased final refined structure may sufficiently demonstrate the unknown structural quantities. On the other hand, insufficient restraint may allow unreasonable geometries that do not allow recognition of the desired quantity. In this case, it may be necessary to test all possible restraint conditions and compare the results of the refinements.

18.3.3.1. Confidence in restraints versus information from diffraction

Primarily in cases of new structures, such as small-ligand- or metal-binding proteins, the refinement may indicate that the expected geometries and applied restraints seem incompatible with evidence from the electron density. Several sources for such discrepancies must be considered for an evaluation of the true geometries or the confidence level of such an evaluation. The quality of the experimental information, such as data resolution and reduction parameters, must be considered. Physical phenomena possibly ignored by the refinement model might include anisotropies of motion and/or electron distribution, or disorder in the crystal. These might lead to systematic deviations in the refined structure that mimic alternate parameterizations. Finally, newly derived parameters should be examined to decide whether the fragments and chemical environments were inappropriate for the refinement problem, or whether errors in fragment structures artifactually distorted the parameterization.

18.3.4. Future perspectives

It seems obvious to seek the best (most accurate) possible parameterization and establish it as a standard (to enable statistical structure comparisons). This does not seem to be a realistic goal for several reasons. Firstly, the parameterization is less a determinant of accuracy than the quality of the data and the method of refinement. Secondly, the quality of existing parameterization and the potential for new environment-dependent parameters improves as more structures are solved and databases grow. Such new parameters can be derived from conformation-dependent statistics (*cis*- and *trans*-proline is an example described above), hydrogen-bonding geometries *etc.* Finally, protein structures are generally solved not to build a statistically optimized protein database, but to discover biophysical functional mechanisms.

The growth of structural databases will improve our understanding of structural properties (Wilson *et al.*, 1998); the highest-resolution protein structures will contribute most to the database, while low-resolution structures will profit most from improved predictive power. Structures that require restrained refinement both draw on the database for refinement parameters and integrity checks, and also contribute to it; a kind of bootstrapping procedure to re-refine deposited structures with iteratively improved parameters is conceivable (if convergent). The consequent removal of parameterization 'signatures' in, *e.g.*, bond and angle parameters seems unlikely to have practical consequences beyond identification of, *e.g.*, catalytically relevant outliers, but qualitative improvements in structure comparison might be revealing in unexpected ways. Such an effort will require adequate computational resources and the deposition of structure factors or, even better, diffraction images.

References

Bansal, M. & Ananthanarayanan, V. S. (1988). *The role of hydroxyproline in collagen folding: conformational energy calculations on oligopeptides containing proline and hydroxyproline. Biopolymers*, **27**, 299–312.

Bricogne, G. (1993). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.

Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Cryst.* D**49**, 24–36.

Bürgi, H.-B. & Dubler-Steudle, K. C. (1988). *Empirical potential energy surfaces relating structure and activation energy. 2. Determination of transition-state structure for the spontaneous hydrolysis of axial tetrahydropyranyl acetals. J. Am. Chem. Soc.* **110**, 7291–7299.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution. Curr. Opin. Struct. Biol.* **7**, 681–688.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci.* **13**, 425–430.

Kidera, A., Matsushima, M. & Go, N. (1994). *Dynamic structure of human lysozyme derived from X-ray crystallography: normal mode refinement. Biophys. Chem.* **50**, 25–31.

Kleywegt, G. J. & Jones, T. A. (1998). *Databases in protein crystallography. Acta Cryst.* D**54**, 1119–1131.

Lamzin, V. S., Dauter, Z. & Wilson, K. S. (1995). *Dictionary of protein stereochemistry. J. Appl. Cryst.* **28**, 338–340.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol.* **231**, 1049–1067.

Longhi, S., Czjzek, M. & Cambillau, C. (1998). *Messages from ultrahigh resolution crystal structures. Curr. Opin. Struct. Biol.* **8**, 730–737.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983). *The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Cryst.* B**39**, 480–490.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.

Priestle, J. P. (1994). *Stereochemical dictionaries for protein structure refinement and model building. Structure*, **2**, 911–913.

Sippl, M. J. (1995). *Knowledge-based potentials for proteins. Curr. Opin. Struct. Biol.* **5**, 229–235.

Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *Occurrence and role of cis peptide bonds in protein structures. J. Mol. Biol.* **214**, 253–260.

Vlassi, M., Dauter, Z., Wilson, K. S. & Kokkinidis, M. (1998). *Structural parameters for proteins derived from the atomic resolution (1.09 Å) structure of a designed variant of the ColE1 ROP protein. Acta Cryst.* D**54**, 1245–1260.

Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S. , Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W., Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R. & Rullmann, J. A. C. (1998). *Who checks the checkers – four validation tools applied to eight atomic resolution structures. J. Mol. Biol.* **276**, 417–436.

International Tables for Crystallography (2012). Vol. F, Chapter 18.4, pp. 485–498.

Chapter 18.4. Refinement at atomic resolution

Z. Dauter, G. N. Murshudov and K. S. Wilson

18.4.1. The atomic model and a definition of atomic resolution

18.4.1.1. The atomic model

X-rays are diffracted by the electrons that are distributed around the atomic nuclei, and the result of an X-ray crystallographic study is the derived three-dimensional electron-density distribution in the unit cell of the crystal. The elegant simplicity and power of X-ray crystallography arise from the fact that molecular structures are composed of discrete atoms that are treated as spherically symmetric in the usual approximation. This property places such strong restraints on the Fourier transform of the crystal structures of small molecules that the phase problem can be solved by knowledge of the amplitudes alone.

Each atom or ion can be described by up to 11 parameters (Table 18.4.1.1).

The first parameter is the scattering-factor amplitude for the chemical nature of the atom in question, and has been computed and tabulated for all atom types [*International Tables for Crystallography*, Volume C (2004)]. Once the chemical identity of the atom is established, this parameter is fixed.

The next three parameters relate to the positional coordinates of the atom with respect to the origin of the unit cell.

If the resolution is high enough, then the number of observed reflections is sufficient to allow six anisotropic atomic displacement parameters to be used to describe the distribution of the atom positions in different unit cells (Fig. 18.4.1.1). Atomic displacement parameters (ADPs) reflect both the thermal vibration of atoms about the mean position as a function of time (dynamic disorder) and the variation of positions between different unit cells of the crystal arising from its imperfection (static disorder). Contributors to the apparent ADP ($U_{\rm atom}$) can be thought of as follows (Murshudov *et al.*, 1999):

$$U_{\rm atom} = U_{\rm crystal} + U_{\rm TLS} + U_{\rm torsion} + U_{\rm bond}, \qquad (18.4.1.1)$$

where $U_{\rm crystal}$ represents the fact that a crystal itself is in general an anisotropic field that will result in the intensity falling off in an anisotropic manner, $U_{\rm TLS}$ represents a translation/libration/screw (TLS), *i.e.* the overall motion of molecules or domains (Schomaker & Trueblood, 1968; Winn *et al.*, 2001), $U_{\rm torsion}$ is the oscillation along torsion angles and $U_{\rm bond}$ is the oscillation along and across bonds. In principle, all these contributors are highly correlated and it is difficult to separate them from one another. Nevertheless, an understanding of how $U_{\rm atom}$ is a sum of these different components makes it possible to apply atomic anisotropy parameters at different resolutions in a different manner. For example, $U_{\rm crystal} + U_{\rm TLS}$ can be applied at any resolution, as their refinement increases the number of parameters by at most five for $U_{\rm crystal}$ and 20 per independent moiety for $U_{\rm TLS}$. In contrast, refinement of the third contributor does pose a problem, as there is strong correlation between different torsion angles. As an alternative, ADPs along the internal degrees of freedom could in principle be refined. The fourth and final contributor, $U_{\rm bond}$, can only be refined at very high resolution. In real applications, $U_{\rm crystal}$ and $U_{\rm TLS}$ are separated for convenient description of the system, but in practice their effects are indistinguishable.

In the special case when the tensor $U_{\rm atom}$ is isotropic, *i.e.*, all non-diagonal elements are equal to zero and all diagonal terms are equal to each other, then the atom itself appears to be isotropic and its ADP can be described using only one parameter, $U_{\rm iso}$.

Thus, for a full description of a crystal structure in which all atoms only occupy a single site, nine parameters per atom must be determined: three positional parameters and six anisotropic ADPs. This assumes that the spherical-atom approximation applies and ignores the so-called deformation density resulting from the non-spherical nature of the outer atomic and molecular orbitals involved in the chemical interactions between the atoms (Coppens, 1997).

For disordered regions or features, where atoms can be distributed over two or more identifiable sites, the occupancy introduces a tenth variable for each atom. In many cases, the fractional occupancies are not all independent, but are rather constant for sets of covalently or hydrogen-bonded atoms or for those in non-overlapping solvent networks. This would apply, for example, to partially occupied ligands or side chains with two conformations.

Table 18.4.1.1
The parameters of an atomic model

Parameter type	Number	Variable or fixed
Atom type	1	Fixed after identification
Positional (x, y, z)	3	Variable, subject to restraints
ADPs:		
isotropic	1	Variable beyond about 2.5 Å
anisotropic	6	Variable beyond about 1.5 Å
Occupancy	1	Variable for visible disorder

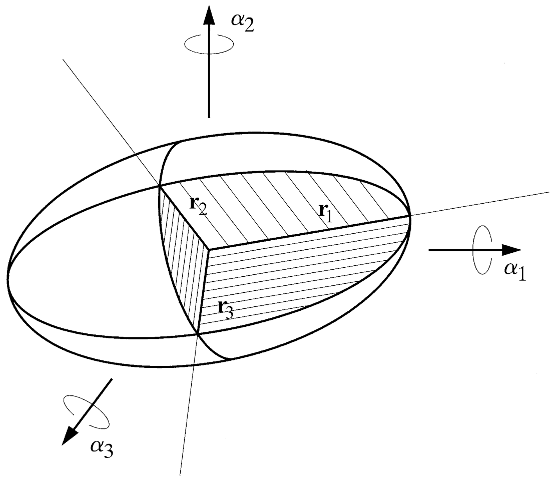


Figure 18.4.1.1
The thermal-ellipsoid model used to represent anisotropic atomic displacement, with major axes indicated. The ellipsoid is drawn with a specified probability of finding an atom inside its contour. Six parameters are necessary to describe the ellipsoid: three represent the dimensions of the major axes and three the orientation of these axes. These six parameters are expressed in terms of a symmetric U tensor and contribute to atomic scattering through the term $\exp[-2\pi^2(U_{11}h^2a^{*2} + U_{22}k^2b^{*2} + U_{33}l^2c^{*2} + 2U_{12}hka^*b^*\cos\gamma^* + 2U_{13}hla^*c^*\cos\beta^* + 2U_{23}klb^*c^*\cos\alpha^*)]$.

Copyright © 2012 International Union of Crystallography

18.4.1.2. *What is 'atomic resolution'?*

Atomicity is the great simplifying feature of crystallography in terms of structure solution and refinement. For a small-molecule structure, accurate X-ray data usually extend to 0.8 Å, and this has three important implications for crystallography.

(1) *Ab initio phasing using direct methods*. Automatic *ab initio* solution of the phase problem depends on the assumption of positivity and atomicity of the electron density. The fact that current *ab initio* methods in the absence of heavy atoms are only effective when meaningful data extend beyond 1.2 Å reinforces the idea that this is a reasonable working criterion for its definition as atomic resolution. In addition, approaches such as solvent flattening and automated map interpretation benefit enormously from such data.

(2) *Resolved atomic peaks in the Fourier maps*. Although some individual peaks can be seen at resolutions beyond ~2.0 Å, they become more fully resolved at around 1.2 Å.

(3) *Refinement of a full anisotropic model*. The number of reflections is sufficient for the minimization of the discrepancy between the experimentally determined amplitudes or intensities of the Bragg reflections and those calculated from the atomic model with up to ten (usually nine) independent parameters per atom. This has been classically achieved by least-squares refinement as described in *International Tables for Crystallography* Volume C, Chapter 8.1 (Prince & Boggs, 2004) or more recently by maximum-likelihood procedures (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997). For small-molecule structures, accurate amplitude data are normally available to around 0.8 Å, giving an observation-to-parameter ratio of about seven for noncentrosymmetric crystals, which allows positional parameters to be determined with an accuracy approaching 0.001 Å. This reflects the high degree of order of such crystals, in which the molecules in the lattice are in a close-packed array. In addition the X-ray data are of high quality, with a high $I/\sigma(I)$ ratio (and hence low merging R value) even in the outer resolution shells.

It is now necessary to define what constitutes 'atomic resolution'. A pragmatic approach has been that data extending to 1.2 Å or better with at least 50% of the intensities in the outer shell being higher than 2σ is the acceptable limit (Sheldrick, 1990; Sheldrick & Schneider, 1997), which means that the statistical problem of refinement is overdetermined. This appears to remain a good working definition for refinement applications and indeed has been put on a more solid theoretical basis (Morris & Bricogne, 2003; Morris *et al.*, 2004). However, for application of direct phasing methods it is advantageous to record even a small fraction of significant reflections beyond this cutoff. These outer shells should be included in the refinement procedure with correct maximum-likelihood weights, but they will not significantly improve the effective resolution.

This is rarely achieved for crystals of macromolecules: as of October 2009 around 1250 out of 52 000 crystal structures in the Protein Data Bank (PDB) had a resolution higher than 1.2 Å compared to 157 out of 13 000 in March 2000. Firstly, the large unit-cell volume leads to an enormous number of reflections for which the average intensity is weak compared to those for small molecules (see Table 9.1.1.1 in Chapter 9.1). Secondly, the intrinsic disorder of the crystals further reduces the intensities at high Bragg angles and usually gives a resolution cutoff which is much less than atomic. Thirdly, the large solvent content leads to substantial decay of crystal quality under exposure to the X-ray beam at room temperature. While the secondary damage (resulting from the migration of ions and radicals produced by the primary absorption event) is largely avoided by vitrification of such crystals, the effect of primary damage has become significant on high intensity beamlines (see Section 9.1.12). The upper resolution limit of the data affects all stages of a crystallographic analysis, but especially restricts the features of the model that can be independently refined (Table 18.4.1.2). Solutions to the problem of refining macromolecular structures with a paucity of experimental data evolved during the 1970s and 1980s with the use of either constraints or restraints on the stereochemistry, based on that of known small molecules. With constraints, the structure is simplified as a set of rigid chemical units (Diamond, 1971; Herzberg & Sussman, 1983), whereas using restraints, the observation-to-parameter ratio is increased by introduction of prior chemical knowledge of bond lengths and angles (Konnert & Hendrickson, 1980).

Table 18.4.1.2
Features which can be seen in the electron density at different resolutions

Disordered regions will not necessarily be visible even at these limiting values. Some features should be included even at lower resolutions, *e.g.* hydrogen atoms at their riding positions can be incorporated at 2.0 Å, but their positions will not be verifiable from the density. The contents of this table should not be taken as dogmatic rules, but as approximate guidelines.

Resolution (Å)	Feature
0.8	Deformation density, *i.e.* deviation from the spherical-atom model
1.0	Hydrogen atoms
1.5	Anisotropic atomic displacement
2.0	Multiple conformations
2.5	Individual isotropic atomic displacement
3.5	Overall temperature factor
4.0	α-Helices and β-sheets
6.0	Domain envelopes

As expected, atoms with different ADPs contribute differently to the diffraction intensities, as discussed by Cruickshank (1999*a*,*b*). The relative contribution of the different atoms to a given reflection depends on the difference between their ADPs $\{\exp[-(B_1 - B_2)s^2]$, where $s = \sin\theta/\lambda\}$. Clearly, if the average ADP of a molecule is small, then the spread will also be narrow, and most atoms will contribute to diffraction over the whole range of resolution. When the mean ADP is large, then the spread of the ADPs will be wide, and fewer atoms will contribute to the high-resolution intensities (Fig. 18.4.1.2).

Three advances in experimental techniques have combined effectively to overcome these problems for an increasing number of well ordered macromolecular crystals, namely the use of high-intensity synchrotron radiation (SR), efficient two-dimensional detectors and cryogenic freezing (discussed in Parts 8, 7 and 10, respectively). These advances mean that there is no longer a sharp division between small-molecule and macromolecular crystallography, but rather a continuum from small through medium-sized structures, such as cyclodextrins and other supramolecules, to proteins. The inherent disorder in the crystal generally increases with the size of the structure, due in part to the increasing solvent content. Thus, it has become tractable to refine a significant number of protein structures at atomic resolution with a full anisotropic model (Dauter, Lamzin & Wilson, 1997; Dauter, 2003). This work of course benefits tremendously from the experience and algorithms of small-molecule crystallography, but does pose special problems of its own. The tech-

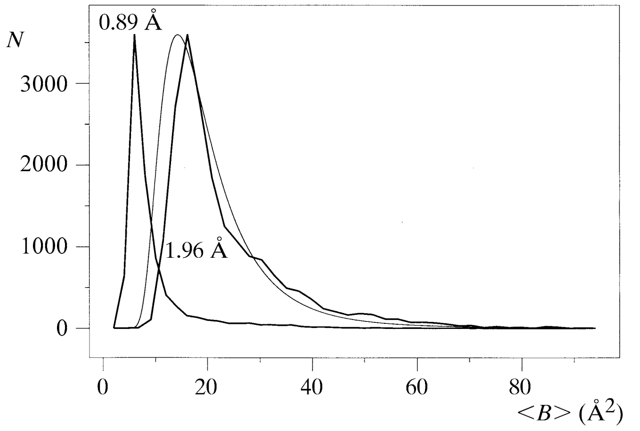


Figure 18.4.1.2
Histograms of B values for a protein structure, *Micrococcus lysodecticus* catalase (Murshudov *et al.*, 1999), for two different crystals which diffracted to different limiting resolutions. For both crystals, the resolution cutoff reflects the real diffraction limit from the sample, and hence its level of order. At 0.89 Å, the mean B value is 8.3 Å^2 and the width of the distribution is small. In contrast, at 1.96 Å, the mean B value is 25.5 Å^2 and the spread is correspondingly large. Thus, for the 0.89 Å crystal, most atoms contribute to the high-resolution terms, whereas for the 1.96 Å crystal, only the atoms with lower B values do so. The thin line shows the theoretical inverse gamma distribution $\mathrm{IG}(B) = (b/2)^{d/2}/\Gamma(d/2)B^{-(d+2)/2}\exp[-b(2B)]$, where b and d are the parameters of the distribution, and Γ is the gamma function. For this figure, the values $b = 2$ and $d = 10$ were chosen, which correspond to a mean B value of 20 Å^2 and σ_B of 11 Å^2. In the gamma distribution, the abscissa was multiplied by $8\pi^2$ to make it comparable with the measured B values. All three histograms were normalized to the same scale.

niques of solving and refining macromolecular structures thus also overlap with those conventionally used for small molecules; a prime example is the use of *SHELXL* (Sheldrick, 2008), which was developed for small structures and has now been extended to treat macromolecules.

18.4.1.3. A theoretical approach to 'atomic resolution'

An alternative and stricter definition of atomic resolution comes from using a measure of the information content of the data. There are a variety of definitions of the information in the data about the postulated model (see, for example, O'Hagan, 1994). A suitable one is the Bayesian definition for quadratic information measure:

$$I_Q(p, F) = \operatorname{tr}(A\{\operatorname{var}(p) - E[\operatorname{var}(p, F)]\}), \qquad (18.4.1.2)$$

where I_Q is the quadratic information measure, p is the vector of parameters, F is the experimental data, var(p) is the variance matrix corresponding to prior knowledge, var(p, F) is the variance matrix corresponding to the posterior distribution (which includes prior knowledge and likelihood), E is the expectation, tr is the trace operator (*i.e.* the sum of the diagonal terms of the matrix) and A is the matrix through which the relative importance of different parameters or combinations of parameters is introduced. For example, if A is the identity matrix, then the information measure is unitary and all parameters are assigned the same weight. If A is the identity matrix for positional parameters and zero for ADPs, then only the information about positional parameters is included. By appropriate choice of A, the information about selected key features, such as the active site, can be estimated.

Equation (18.4.1.2) shows how much the experiment reduces the uncertainty in given parameters. Prior knowledge is usually taken to be information about bond lengths, bond angles and other chemical features of the molecule, known before the experiment has been carried out. In the case of an experiment designed to provide information about the ligated protein or mutant, when information about differences between two (or more) different states is needed, the prior knowledge can be thought of in a different way – as knowledge about the native protein.

Unfortunately, there are problems in applying equation (18.4.1.2). Firstly, careful analysis of the prior knowledge and its variance is essential. The target values used at present, or more properly the distributions for these values, need to be re-evaluated. Another problem concerns the integration required to compute the expectation value (E). Nevertheless, the equation provides some idea of how much information about a postulated model can be extracted from a given experiment.

This alternative definition of atomic resolution assumes that the second term of equation (18.4.1.2) for positional parameters is sufficiently close to zero for most atoms to be resolved from all their neighbours. Defining atomic resolution using this information measure reflects the importance of both the quality and quantity of the data [through the posterior var(p, F)]. In addition, data may come from more than one crystal, in which case the information will be correspondingly increased. There may be additional data from mutant and/or complexed protein crystals, where, again, the information measure will be increased and, moreover, the differences between different states can be analysed. The effect of redundancy of different crystals of the same molecule(s) in different space groups is to reduce the limit of data necessary for achieving atomic resolution, which is equivalent to the advantage of noncrystallographic averaging.

Thus, in practice, while it would be ideal to develop the strict application of equation (18.4.1.2), for the present it is necessary to rely on the pragmatic approach in Section 18.4.1.2.

18.4.2. Data

The quality of the refined model relies finally on that of the available experimental data. Data collection has been covered extensively in Chapter 9.1 and will not be discussed here.

18.4.2.1. Data quality

As can be seen from equation (18.4.1.2), the measure of information about all or part of the crystal contents depends strongly on the quality and quantity of the data. Of course, before the experiment is carried out some questions should be answered. Firstly, what is the aim of the experiment? Secondly, what is the cost of the experiment and what are the available resources? With modern techniques, if SR is used with an efficient detector, the cost of the experiment for different resolutions does not vary greatly (provided that a suitable quality crystal is available). In practice, the apparent increase in cost to attain high-resolution data will generally make solving the phase problem both easier and faster. A full analysis at atomic resolution provides a wealth of additional structural detail which may shed light on the subtleties of the protein's chemistry not seen at lower resolution. However, this may require some considerable time and effort, and is an area where development of more automated approaches would be beneficial. In contrast, low-resolution data can make it difficult to answer not only the question currently being asked, but can also necessitate further experiments to address other problems that arise.

While the information content of the data appears to depend quantitatively on the nominal resolution, in fact it is dependent on the data quality throughout the resolution range, and both high- and low-resolution completeness and their statistical significance affect the information content of the data and derived model. High-intensity low-resolution terms remain important for refinement at atomic resolution, as they define the contrast in the density maps between solvent and protein, and because their omission biases the refinement, especially that of parameters such as the ADPs. To judge the effective resolution of the diffraction data set, the concept of 'optical resolution', which can be estimated from the shape of the origin peak in the Patterson synthesis, may be very useful (Vaguine *et al.*, 1999).

The rejection of low-intensity observations will also introduce bias. In particular, all the maps calculated for visual or computer inspection by Fourier transformation are diminished in quality by omission of any terms, but are especially affected by omission of strong low-resolution data. This is particularly true in the early stages of structure solution, where low-resolution data can be vital. Although most phase-improvement algorithms rely on relations between all reflections, terms involving low-resolution reflections will be large, will be involved in many relations and will play a dominant role. Hence, omission of these terms will severely degrade the power of these methods, which may indeed converge to solutions that have nothing whatsoever to do with the real structure.

18.4.2.2. Anisotropic scaling

The intensity data from a crystal may display anisotropy, *i.e.*, the intensity fall-off with resolution will vary with direction, and may be much higher along one crystal axis than along another. If the structure is to be refined with an isotropic atomic model (either because there are insufficient data or the programs used cannot handle anisotropic parameters), then the fall-off of the calculated F^2 values will, of necessity, also be isotropic. In this situation, an improved agreement between observed and calculated F^2 values can be obtained either by using anisotropic scaling during data reduction to the expected Wilson distribution of intensities, or by including a maximum of six overall anisotropic parameters during refinement. This will result in an isotropic set of F^2 values. For crystals with a high degree of anisotropy in the experimental data, this can lead to a substantial drop of several per cent in R and $R_{\rm free}$ (Sheriff & Hendrickson, 1987; Murshudov *et al.*, 1998).

This ambiguity effectively disappears with use of an anisotropic atomic model. The individual ADPs, including contributions from both static and thermal disorder, take up relative individual displacements, but also the overall anisotropy of the experimental F^2 values. The significance of the overall anisotropy is a point of some contention, and its physical meaning is not clear. It may represent asymmetric crystal imperfection or anisotropic overall displacement of molecules in the lattice related to TLS parameters. Refinement of TLS parameters, which can be performed using, for example, *RESTRAIN* (Driessen *et al.*, 1989) *REFMAC* (Winn *et al.*, 2001) or *PHENIX.REFINE* (Terwilliger *et al.*, 2008), removes the overall crystal contribution to the ADPs.

It is important that at least the intensity or amplitude data be deposited in the PDB as measured, without any anisotropic correction being applied. For the refined model, complete information on any overall anisotropic and TLS modelling must be explicitly included, as well as the individual atomic ADPs.

18.4.3. Computational algorithms and strategies

18.4.3.1. Classical least-squares refinement of small molecules

The principles of the least-squares method of minimization are described in *International Tables for Crystallography* Volume C (2004). Least squares involves the construction of an order $N \times N$ normal matrix, where N is the number of parameters, representing a system of least-squares equations, whose solution provides estimates of adjustments to the current atomic parameters. The problem is nonlinear and the matrix construction and solution must be iterated until convergence is achieved. In addition, inversion of the matrix at convergence provides an approximation to the standard uncertainties for each individual parameter refined according to the Cramer–Rao inequality (Stuart *et al.*, 1999). Indeed, this is the only method available so far that gives such estimates properly.

However, even for small molecules there may be some disordered regions which will require the imposition of restraints, as is the case for macromolecules (see below), and the presence of such restraints means that the error estimates no longer reflect the information from the X-ray data alone. If the problem of how restraints affect the error estimates could be resolved, then inversion of the matrix corresponding to the second derivative of the posterior distribution would provide standard uncertainties incorporating both the prior knowledge, such as the restraints, and the experimental data. Equation (18.4.1.2) for information measure could then be applied, but this requires further development. For small structures, the speed and memory of modern computers have reduced the requirements for such calculations to the level of seconds, and the computational requirements form a trivial part of the structure analysis. Recent developments in the application of fast Fourier transform methods to normal matrix–vector multiplication (Strokopytov, 2008) and fast information matrix evaluation (Steiner *et al.*, 2003) suggest that fast calculations for macromolecular structures may be available in the foreseeable future.

18.4.3.2. Least-squares refinement of large structures

The size of the computational problem increases dramatically with the size of the unit cell, as the number of terms in the matrix increases with the square of the number of parameters. Furthermore, construction of each element depends on the number of reflections. For macromolecular structures, computation of a full matrix is at present prohibitively expensive in terms of CPU time and memory. A variety of simplifying approaches have been developed, but all suffer from a poorer estimate of the standard uncertainties and from a more limited range and speed of convergence.

The first is the block-matrix approach, where instead of the full matrix, only square blocks along the matrix diagonal are constructed, involving groups of parameters that are expected to be correlated. The correlation between parameters belonging to different blocks is therefore neglected completely. In this way, the whole least-squares minimization is split into a set of smaller independent units. In principle this leads to the same solution, but more slowly and with less precise error estimates. Nevertheless, block-matrix approaches remain essential for tractable matrix inversion for macromolecular structures.

A further simplification involves the conjugate-gradient method or the diagonal approximation to the normal matrix (the second derivative of minus the log of the likelihood function in the case of maximum likelihood), which essentially ignores all off-diagonal terms of the least-squares matrix. For the conjugate-gradient approach, all diagonal terms of the matrix are equal. However, the range and speed of convergence are substantially reduced, and standard uncertainties can no longer be estimated directly by matrix inversion.

18.4.3.3. Fast Fourier transform

Conventional least-squares programs use the structure-factor equation and associated derivatives, with the summation extending over all atoms and all reflections. This is immensely slow in computational terms for large structures, but it has the advantage of providing precise values.

An alternative procedure, where the computer time is reduced from being proportional to N^2 to $N \log N$, involves the use of fast Fourier algorithms for the computation of structure factors and derivatives (Ten Eyck, 1973, 1977; Agarwal, 1978). This can involve some interpolation and the limitation of the volume of electron-density maps to which individual atoms contribute. Such algorithms have been exploited extensively in macromolecular refinement programs such as *PROLSQ* (Konnert & Hendrickson, 1980), *XPLOR* (Brünger, 1992*b*), *TNT* (Tronrud, 1997), *RESTRAIN* (Driessen *et al.*, 1989), *REFMAC* (Murshudov *et al.*, 1997), *CNS* (Brünger *et al.*, 1998) and *PHENIX.REFINE* (Terwilliger *et al.*, 2008), but have largely been restricted to the diagonal approximation. *XPLOR* and *CNS* use the conjugate-gradient method, which relies only on the first derivatives and ignores the second derivatives. In all other programs, the diagonal approximation is used for the second-derivative matrix.

18.4.3.4. Maximum likelihood

This provides a statistically sounder alternative to least squares, especially in the early stages of refinement when the model lies far from the minimum. This approach increases the radius of convergence, takes into account experimental uncertainties, and in the final stages gives results similar to least squares but with improved weights (Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). The maximum-likelihood approach has been extended to allow refinement of a full atomic anisotropic model while retaining the use of fast Fourier algorithms (Murshudov *et al.*, 1999). A remaining limitation is the use of the diagonal approximation, which prevents the computation of standard uncertainties of individual parameters. Algorithms that will alleviate this limitation can be foreseen, and they are expected to be implemented in the future.

18.4.3.5. Twinning

A non-negligible fraction of protein crystals turn out to be merohedrally twinned (Lebedev *et al.*, 2006), which requires special treatment of the diffraction data and proper treatment of this phenomenon during structure solution and more especially refinement (Yeates, 1997; Chapters 18.11 and 18.12). The twinning problem can be approached in two distinctly different ways. In the first, the data are explicitly 'detwinned' to produce an adjusted set of amplitudes [for example using the UCLA detwinning server (http://nihserver.mbi.ucla.edu/Twinning/) or the *DETWIN* program (Collaborative Computational Project, Number 4, 1994)], which are then subjected to conventional refinement. However, detwinning can only be successfully applied if the twinning fraction is not too high, since the error in the resulting amplitudes increases as the fraction approaches 50%. A second, and certainly preferred, approach is to include the twinning fraction as a variable during refinement. This has been implemented in *SHELXL*, *PHENIX.REFINE* and *REFMAC*, and is widely used. In *CNS*, it is possible to set the twin fraction, but not to refine it. This is of special relevance for atomic resolution structures, as even a small degree of twinning will have a significant effect on the interpretability of fine features.

18.4.3.6. Computer power

There are no longer any restrictions on the full-matrix refinement of small-molecule crystal structures. However, the large size of the matrix, which increases as N^2, where N is the number of parameters, means that for macromolecules with thousands of independent atoms this approach is intractable with the computing resources normally available to the crystallographer. By extrapolating the progress in computing power experienced in recent years, it can be envisaged that the limitations will disappear during the next decade, as those for small structures have disappeared since the 1960s. Indeed, the advances in the speed of CPUs, computer memory and disk capacity continue to transform the field.

18.4.4. Computational options and tactics

18.4.4.1. Use of F (amplitudes) or F^2 (intensities)

The X-ray experiment provides two-dimensional diffraction images. These are transformed to integrated but unscaled data, which are transformed to Bragg reflection intensities that are subsequently transformed to structure-factor amplitudes. At each transformation some assumptions are used, and the results will depend on their validity. Invalid assumptions will introduce bias toward these assumptions into the resulting data. Ideally, refinement (or estimation of parameters) should be against data that are as close as possible to the experimental observations, eliminating at least some of the invalid assumptions. Extrapolating this to the extreme, refinement should use the images as observable data, but this poses several severe problems, depending on data quantity and the lack of an appropriate statistical model.

Alternatively, the transformation of data could be improved by revising the assumptions. The intensities are closer to the real experiment than are the structure-factor amplitudes, and use of intensities would reduce the bias. However, there are some difficulties in the implementation of intensity-based likelihood refinement (Pannu & Read, 1996).

Gaussian approximation to intensity-based likelihood (Murshudov *et al.*, 1997) would avoid these difficulties, since a Gaussian distribution of error can be assumed in the intensities but not the amplitudes. However, errors in intensities may not only be the result of counting statistics, but may have additional contributions from factors such as crystal disorder and motion of the molecules in the lattice during data collection.

Nevertheless, the problem of how to treat weak reflections remains. Some of the measured intensities will be negative, as a result of statistical errors of observation, and the proportion of such measurements will be relatively large for weakly diffracting macromolecular structures, especially at atomic resolution. This is less important for intensity-based likelihood than for the amplitude-based approach. French & Wilson (1978) have given a Bayesian approach for the derivation of structure-factor amplitudes from intensities using Wilson's distribution (Wilson, 1942) as a prior, but there is room for improvement in this approach. Firstly, the Wilson distribution could be upgraded using the scaling techniques suggested by Blessing (1997) and Cowtan &

Main (1998), and secondly, information about effects such as pseudosymmetry could be exploited.

Another argument for the use of intensities rather than amplitudes is relevant to least squares, where the derivative for amplitude-based refinement with respect to $F_{\rm calc}$ is singular when $F_{\rm calc}$ is equal to zero (Schwarzenbach *et al.*, 1995). This is not the case for intensity-based least squares. In applying maximum likelihood, this problem does not arise (Pannu & Read, 1996; Murshudov *et al.*, 1997).

Finally, while there may be some advantages in refining against intensities, Fourier syntheses always require structure-factor amplitudes.

18.4.4.2. Restraints on coordinates and ADPs

For a good small-molecule crystal the experimental X-ray data extend to ~0.8 Å spacing and the structure can be refined against the X-ray data alone. The resulting accuracy of the atomic coordinates will generally be better than 0.01 Å. However, even for small-molecule structures, disordered regions require the imposition of stereochemical restraints (or constraints) if the chemical integrity is to be preserved and the ADPs are to be realistic.

The typical situation for protein crystals is quite different, with atomic resolution being the exception rather than the rule. Thus, for proteins the geometry of the atomic model needs to be restrained, both in terms of geometry and ADPs. The geometric target values have been established from a set of amino-acid and small-peptide structures (see Chapter 18.3 by Engh & Huber), for which the bond-length r.m.s.d. is about 0.02 Å. In the present context, we restrict the discussion to bond lengths, but this is representative of the other restraints. Clearly, the relative contribution of the X-ray data and the restraints on the final parameters varies as a function of resolution. The restraints dominate at low resolution, while by the time 0.8 Å spacing is achieved, the restraints will be essentially irrelevant for well ordered regions. The imposition of bond-length restraints with target deviations of ~0.02 Å means that the distribution of bond lengths in the final model will be of the same order, independent of the resolution of the X-ray data. However, this must not be taken to imply that the accuracy of the atomic parameters is invariant with the overall resolution and, more importantly, the atomic displacement. To be explicit, the accuracy of the atomic positions decreases (1) as the resolution becomes worse (*i.e.*, the number of X-ray observations decreases) and (2) as the ADPs become larger. While this should be obvious to the practicing crystallographer, it may not be so apparent to the less expert user of the PDB (Wlodawer *et al.*, 2008).

Jaskolski *et al.* (2007) analysed ten structures from the PDB refined at ultra-high (better than 0.8 Å) resolution to investigate appropriate geometrical restraints. They confirmed the general correctness of values in the Engh & Huber dictionary and showed that the mean observed deviations in these ten structures from the target bond lengths were indeed roughly 0.02 Å. They therefore postulated that 0.02 Å was an appropriate value to use in applying stereochemical restraints to protein structures in general. There has been some dispute about this value (Tickle, 2007) but we believe it to be appropriate. A more detailed analysis of this issue has been performed, suggesting that target values differ depending on the structural context (Karplus *et al.*, 2008); this may lead to some fine adjustments in the target dictionary, as predicted earlier (EU 3-D Validation Network, 1998).

In analogy to the geometrical restraints based on the Engh & Huber dictionary, anisotropic ADP restraints were established in the *SHELX* program suite (Sheldrick, 2008). For example, they prevent atoms from becoming unrealistically anisotropic and restrain the shapes of the ellipsoids of bonded atoms to be not too dissimilar. Riding hydrogen atoms are assigned constrained isotropic ADPs based on those of the parent atoms.

A more theoretical justification for use of restraints is that refinement can be considered as Bayesian estimation. From this point of view, all available and usable prior knowledge should be exploited, as it should not harm the parameter estimation during refinement. Bayesian estimation shows asymptotic behaviour (Box & Tiao, 1973), *i.e.*, when the number of observations becomes large, the experimental data override the prior knowledge. In this sense, the purpose of the experiment is to enhance our knowledge about the molecule, and the procedure should be cumulative, *i.e.*, the result of the old experiment should serve as prior knowledge for the design and treatment of new experiments (Box & Tiao, 1973; Stuart *et al.*, 1999; O'Hagan, 1994). However, there are problems in using restraints. For example, the probability distribution reflecting the degree of belief in the restraints is not good enough. Use of a Gaussian approximation to distributions of distances, angles and other geometric properties has not been justified. Firstly, the distribution of geometric parameters depends strongly on ADPs. Secondly, different geometric parameters are correlated. Thirdly, many geometric parameters (*e.g.* bond angles, torsion angles) are dependent on the conformation, configuration and environment of the molecule in question (Karplus *et al.*, 2008; Gelbin *et al.*, 1996). This problem should be the subject of further investigation.

In summary, the atomic resolution structures to date confirm that a mean deviation of bond lengths from target values of 0.02 Å (and comparable values for other restraint types) is appropriate, but may be subject to minor adjustments. These levels of restraint should be applied at all resolutions: the stereochemistry should be neither over- nor under-restrained.

18.4.4.3. Partial occupancy

It may be necessary to refine one additional parameter, the occupancy factor of an atomic site, for structures possessing regions that are spatially or temporally disordered, with some atoms lying in more than one discrete site. The sum of the occupancies for alternative individual sites of a protein atom must be 1.0.

For macromolecules, the occupancy factor is important in several situations, including the following:

(1) when a protein or ligand atom is present in all molecules in the lattice, but can lie in more than one position due to alternative conformations;
(2) for the solvent region, where there may be overlapping and mutually exclusive solvent networks;
(3) when ligand-binding sites are only partially occupied due to weak binding constants, and the structures represent a mixture of native enzyme with associated solvent and the complex structure;
(4) when there is a mixture of protein residues in the crystal, due to inhomogeneity of the sample arising from polymorphism, a mixture of mutant and wild-type protein, or other causes.

Unfortunately, the occupancy parameter is highly correlated with the ADP, and it is difficult to model these two parameters at resolutions less than atomic. Even at atomic resolution, it can

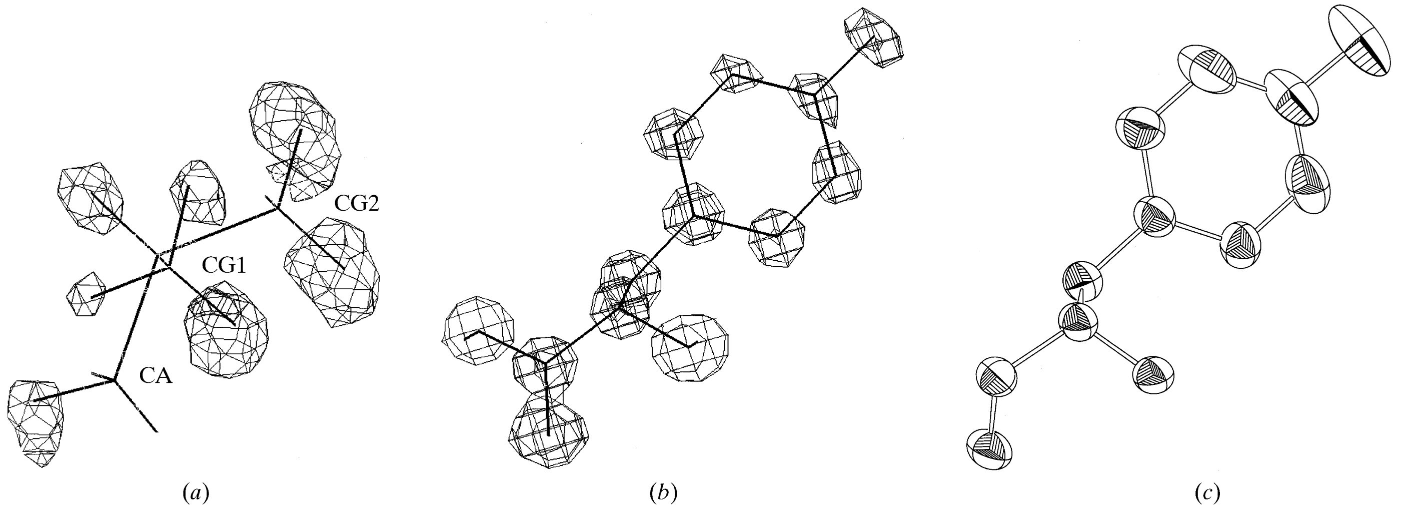


Figure 18.4.5.1
(*a*), (*b*) Representative electron-density maps for the refinement of *Clostridium acidurici* ferredoxin at 0.94 Å resolution (Dauter, Wilson *et al.*, 1997). (*a*) The density for hydrogen atoms (at 3σ) omitted from the structure-factor calculation for Val42. (*b*) The $(2F_o - F_c)$ density for Tyr30, contoured at 3σ. (*c*) The thermal ellipsoids corresponding to (*b*), drawn at the 33% probability level using *ORTEP*II (Johnson, 1976). There is a clear correlation between the density in (*b*) and the ellipsoids in (*c*), showing increased displacement towards the end of the side chain, particularly in the plane of the phenyl ring.

prove difficult to refine the occupancy satisfactorily with statistical certainty.

18.4.4.4. Validation of extra parameters during the refinement process

The introduction of additional parameters into the model always results in a reduction in the least-squares or maximum-likelihood residual – in crystallographic terms, the *R* factor. However, the statistical significance of this reduction is not always clear, since this simultaneously reduces the observation-to-parameter ratio. It is therefore important to validate the significance of the introduction of further parameters into the model on a statistical basis.

Brünger (1992*a*) introduced the concept of statistical cross validation to evaluate the significance of introducing extra features into the atomic model. For this, a small and randomly distributed subset of the experimental observations is excluded from the refinement procedure, and the residual against this subset of reflections is termed $R_{\rm free}$. It is generally sufficient to include about 1000 reflections in the $R_{\rm free}$ subset; further increase in this number provides little, if any, statistical advantage but diminishes the power of the minimization procedure. For atomic resolution structures, cross validation is important in establishing whether the introduction of an additional type of feature to the model (with its associated increase in parameters) is justified. There are two limitations to this. Firstly, if $R_{\rm free}$ shows zero or minimal decrease compared to that in the *R* factor, the significance remains unclear. Secondly, the introduction of individual features, for example the partial occupancy of five water molecules, can provide only a very small change in $R_{\rm free}$, which will be impossible to substantiate. To recapitulate, at atomic resolution the prime use of cross validation is in establishing protocols with regard to extended sets of parameter types. The sets thus defined will depend on the quality of the data.

In the final analysis, validation of individual features depends on the electron density, and Fourier maps must be judiciously inspected. Nevertheless, this remains a somewhat subjective approach and is in practice intractable for extensive sets of parameters, such as the occupancies and ADPs of all solvent sites. For the latter, automated procedures, which are being developed at present, are an absolute necessity, but they may not be optimal in the final stages of structure analysis, and visual inspection of the model and density is often needed.

The problems of limited data and reparameterization of the model remain. At high resolution, reparameterization means having the same number of atoms, but changing the number of parameters to increase their statistical significance, for example switching from an anisotropic to an isotropic atomic model or *vice versa*. In contrast, when reparameterization is applied at low resolution, this usually involves constraints, *i.e.*, a reduction in the number of independent atoms, but this is not an ideal procedure, as real chemical entities of the model are sacrificed. Reducing the number of independent atoms will inevitably result in disagreement between the experiment and model, which in turn will affect the precision of other parameters. It would be more appropriate to reduce the number of parameters without sacrificing the number of atoms, for example by describing the model in torsion-angle space. Water poses a particular problem, as at low as well as at high resolution the water molecules cannot all be described as discrete atoms. Algorithms are needed to describe them as a continuous model with only a few parameters. In the simplest model, the solvent can be described as a constant electron density.

18.4.5. Features in the refined model

All features of the refined model are more accurately defined if the data extend to higher resolution (Fig. 18.4.5.1). In this section, those features that are especially enhanced in an atomic resolution analysis are described. Introduction of an additional feature to the model should be assessed by the use of cross- or self-validation tools: only then can the significance of the parameters added to the model be substantiated.

18.4.5.1. Hydrogen atoms

Hydrogen atoms possess only a single electron and therefore have low electron density and are relatively poorly defined in X-ray studies. They play central roles in the function of proteins, but at the traditional resolution limits of macromolecular structure analyses their positions can only be inferred rather than defined from the experimental data. Indeed, even at a resolution of 2.5 Å, hydrogen atoms should be included in the refined model, as their exclusion biases the position of the heavier atoms, but with their 'riding' positions fixed by those of the parent atoms.

As for small structures, independent refinement of hydrogen-atom positions is not always warranted, even by atomic resolution data, and hydrogen atoms are rather attached as riding rigidly on the positions of the parent atoms. Nevertheless, atomic resolution data allow the experimental confirmation of the positions of many of the hydrogen atoms in the electron-density maps, as they do account for one-sixth of the diffracting power of a carbon atom. Inspection of the maps can in principle allow the identification of (1) the presence or absence of hydrogen atoms on key residues, such as histidine, aspartate and glutamate or on ligands, and (2) the correct location of hydrogen atoms where more than one position is possible, such as in the hydroxyl groups of serine, threonine or tyrosine.

The correct placement of hydrogen atoms riding on their parent atoms involves computation of the appropriate position after each cycle of refinement. This is done automatically by programs such as *SHELXL*, *REFMAC* or *PHENIX.REFINE*. For rigid groups such as the NH amide, aromatic rings, $-CH_2-$ or $=CH-$, the position is accurately defined by the bonding scheme. For groups such as methyl CH_3 or OH, the position is not absolutely defined, and the software is required to make judgmental decisions. For example, *SHELXL* offers the opportunity to inspect the maximum density on a circular Fourier synthesis for optimal positioning. The bond length is fixed according to results from a small-molecule database. The location of hydrogen atoms on polar atoms can be assisted by software that analyses the local hydrogen-bonding networks; this involves maximization of the hydrogen-bonding potential of the relevant groups. Sheldrick advocates assigning ADP values to riding hydrogen atoms of 1.2 times that of the parent atom, or 1.5 times in the case of methyl and similar entities.

18.4.5.2. *Anisotropic atomic displacement parameters*

Refinement of an isotropic model involves four independent parameters per atom, three positional and one isotropic ADP. In contrast, an anisotropic model requires nine parameters per atom, with the anisotropic atomic displacement described by an ellipsoid represented by six parameters. At 1 Å resolution, the data certainly justify an anisotropic atomic model. Extension of the model from isotropic to anisotropic should generally result in a reduction in the R factor of the order of 5–6% and a comparable drop in $R_{\rm free}$. As a consequence of the diminution of the observable-to-parameter ratio, the R factor at all resolutions will drop by a similar amount; however, $R_{\rm free}$ will not. Experience shows that at 2 Å or less there is no drop in $R_{\rm free}$, and an anisotropic model is totally unsupported by the data. At intermediate resolutions, the result depends on the data quality and completeness. At lower resolution, to account for anisotropy of the atoms, the overall motion of molecules or domains can be refined using translation/libration/screw (TLS) parameters (Schomaker & Trueblood, 1968).

Until the end of the 1990s, anisotropic ADPs had only been handled by programs originally developed for small-molecule analysis, which use conventional algebraic computations of the calculated structure-factor amplitudes, *SHELXL* being a prime example. A limitation of this approach is the substantial computation time required. The use of fast-Fourier-transform algorithms for the structure-factor calculation leads to a significant saving in time (Murshudov *et al.*, 1999). Anisotropic modelling of the individual ADPs is essential if the thermal vibration is to be analysed in terms of coordinated motion of the whole molecule or of domains (Schomaker & Trueblood, 1968). Painter & Merritt (2006) have provided a means of analysing refined models to suggest appropriate TLS groupings.

18.4.5.3. *Alternative conformations*

Proteins are not rigid units with a single allowed conformation. *In vivo* they spontaneously fold from a linear sequence of amino acids to provide a three-dimensional phenotype that may exhibit substantial flexibility, which can play a central role in biological function, for example in the induced fit of an enzyme by a substrate or in allosteric conformational changes. Flexibility is reflected in the nature of the protein crystals, in particular the presence of regions of disordered solvent between neighbouring macromolecules in the lattice (see Section 18.4.5.6).

The structure tends to be highly ordered at the core of the protein, or more properly, at the core of the individual domains. Atoms in these regions in the most ordered protein crystals have ADP values comparable to those of small molecules, reflecting the fact that they are, in essence, closely packed by surrounding protein. In general, as one moves towards the surface of the protein, the situation becomes increasingly fluid. Side chains and even limited stretches of the main chain may show two (or multiple) conformations. These may be significant for the biological function of the protein.

The ability to model the alternative conformations is highly resolution dependent. At atomic resolution, the occupancy of two alternative but well defined conformations can be refined to an accuracy of about 5%, thus second conformations can be seen, provided that their occupancy is about 10% or higher. The limited number of proteins for which atomic resolution structures are available suggest that up to 20% of the 'ordered residues' show multiple conformations. This confers even further complexity on the description of the protein model. A constraint can be imposed on residues with multiple conformations: namely that the sum of all the alternatives must be 1.0. Protein regions (whether they are side- or main-chain regions) with alternative conformations and partial occupancy can form clusters in the unit cell with complementary occupancy. This often coincides with alternative sets of solvent sites, which should also be refined with complementary occupancies.

The atoms in two alternative conformations occupy independent and discrete sites in the lattice, about which each vibrates. However, if the spacing between two sites is small and the vibration of each is large, then it becomes impossible to differentiate a single site with high anisotropy from two separate sites. There is no absolute rule for such cases: programs such as *SHELXL* place an upper limit on the anisotropy and then suggest splitting the atom over two sites. Some regions can show even higher levels of disorder, with no electron density being visible for their constituent atoms. Such fully disordered regions do not contribute to the diffraction at high resolution, and the definition of their location will not be improved with atomic resolution data.

18.4.5.4. Ordered solvent water

A protein crystal typically contains ~50% aqueous solvent. This is roughly divided into two separate zones. The first is a set of highly ordered sites close to the surface of the protein. The second, lying remote from the protein surface, is essentially composed of fluid water, with no order between different unit cells.

At room temperature, the solvent sites around the surface are assumed to be in dynamic equilibrium with the surrounding fluid, as for a protein in solution. Nevertheless, the observation of apparently ordered solvent sites on the surface indicates that these are occupied most of the time. The waters are organized in hydrogen-bonded networks, both to the protein and with one another. The most highly ordered water sites lie in the first solvent shell, where at least one contact is made directly to the protein. For the second and subsequent shells the degree of order diminishes: such shells form an intermediate grey level between the ordered protein and the totally disordered fluid. Indeed, the flexible residues on the surface form part of the continuum between a solid and liquid phase.

In the ordered region, the solvent structure can be modelled by discrete sites whose positional parameters and ADPs can be refined. For sites with low ADPs, the refinement is stable and their behaviour is well defined. As the ADPs increase, or more likely the associated occupancy in a particular site falls, the behaviour deteriorates, until finally the existence of the site becomes dubious. There is no hard cutoff for the reality of a weak solvent site. However, the number and significance of solvent sites are increased by atomic resolution data. Despite the fact that the waters contribute only weakly to the high-resolution terms, the improved accuracy of the rest of the structure and the reduction of noise due to the high resolution mean that their positions become better defined.

Indeed, the occupancy of some solvent sites can be refined if the resolution is sufficient, or at least their fractional occupancy can be estimated and kept fixed (Walsh *et al.*, 1998). This leads to the possibility of defining overlapping water networks with alternative hydrogen-bonding schemes. This can be a most time-consuming step in atomic resolution refinement, and a trade-off finally has to be made between the relevance of any improvement in the model and the time spent.

18.4.5.5. Automatic location of water sites

The protein itself has a clearly defined chemical structure, and the number of atoms to be positioned and how they are bonded to one another are known at the start of model building. The solvent region is in marked contrast to this, as the number of ordered water sites is not known *a priori*, and the distances between them are less well defined, their occupancy is uncertain, and there may be overlapping networks of partially occupied solvent sites. Those of low occupancy lie at the level of significance of the Fourier maps.

Selection of partially occupied solvent sites poses a most cumbersome problem in the modelling over and above that of the macromolecule itself, and can be highly subjective and very time consuming. Improved resolution of the data reveals additional weak or partially occupied solvent sites, which generally do not behave well during refinement. Water atoms modelled into relatively weak peaks in electron density tend to drift out of the density during refinement due to the inaccurate gradients that define their positions.

Given the huge number of water sites in question, automatic and at least semi-objective protocols are required. Several procedures have been developed for the automated identification of water sites during refinement [*inter alia ARP* (Lamzin & Wilson, 1997) and *SHELXL* (Sheldrick & Schneider, 1997)] and others allow selective inspection of such sites using graphics [such as *O* (Jones *et al.*, 1991) and *COOT* (Emsley & Cowtan, 2004)]. These depend on a combination of peak height in the density map and geometric considerations. However, these programs are currently optimized for structures at more typical resolutions, and future efforts could be made to adapt them for atomic resolution structures with overlapping water networks and other high levels of detail.

18.4.5.6. Bulk solvent and the low-resolution reflections

As stated in Section 18.4.5.3 and first reviewed by Matthews (1968) and more recently by Andersson & Hovmöller (1998), macromolecular crystals contain substantial regions of totally disordered, or bulk, aqueous solvent, in addition to those solvent molecules bound to the surface. The average electron density of the crystal volume occupied by protein is 1.35 g cm^{-3} (according to Matthews) or 1.22 g cm^{-3} (according to Andersson & Hovmöller), while that of water is 1.0 g cm^{-3}. This is because the atoms are more closely packed within the protein, as they are connected by covalent bonds, while in solvent regions they form sets of hydrogen-bonded networks.

To model both the solvent and protein regions of the crystal appropriately, it is necessary to have a satisfactory representation of the bulk solvent. The high *R* factors generally observed for most proteins for the low-resolution shells are in part symptomatic of the poor modelling of this feature or of systematic errors in the recording of the intensities of the low-angle reflections. For atomic resolution structures, the *R* factor can fall to values as low as 6–7% around 3–5 Å resolution. However, in lower-resolution shells it then rises steadily, often reaching values in the range of 20–40% below 10 Å. These observations indicate serious deficiencies in our current models or data.

The poorest approach is to ignore bulk solvent and assign zero electron density to those regions where there are no discrete atomic sites, as this leads to a severe discontinuum. An improved approach is to assign a constant value of the electron density to all points of the Fourier transform that are not covered by the discrete, ordered sites. This provides substantial reduction in the *R* factor for low-resolution shells of the order of 10% and requires the introduction of only one extra parameter to the least-squares minimization. An improvement of this simplistic model is the introduction of a second parameter, $B_{\rm sol}$, described by

$$\text{scale} = k_0 \exp(-B_0 s^2/4)[1 - k_{\rm sol} \exp(-B_{\rm sol} s^2/4)], \qquad (18.4.5.1)$$

where k_0 and B_0 are the scale factors for the protein, and $k_{\rm sol}$ and $B_{\rm sol}$ are the equivalent parameters for the bulk solvent (Tronrud, 1997). In effect, this provides a resolution-dependent smoothing of the interface contribution, rather than an overall term applied equally to all the data. The physical basis of this is discussed by Tronrud and implemented in several programs, for example *SHELXL*, *REFMAC* and *PHENIX.REFINE* (Fig. 18.4.5.2).

Another approach (Jiang & Brunger, 1994) to account for the effect of solvent is as follows: (1) calculate the mask covering the atoms in the crystal; (2) include a constant value for the electron density in the region not covered by the mask; (3) calculate structure factors from the solvent region; and (4) compute a total structure by applying the appropriate scale to the solvent contribution:

$$F_{\text{total}} = F_{\text{prot}} + k_m \exp(-B_m s^2/4) F_{\text{mask}}, \qquad (18.4.5.2)$$

where k_m and B_m are the scale and temperature factor for the solvent contribution, and F_{total}, F_{prot} and F_{mask} are the complex structure factors corresponding to the combined contribution from the protein and solvent region. The scale and B values are usually refined iteratively to find the best match to the observed structure factors, and are implemented in *CNS*, *REFMAC* and *PHENIX.REFINE*.

Nevertheless, there remain severe problems in the modelling of the interface. The border between the two regions is not abrupt, as there is a smooth and continuous change from the region with fully occupied, discrete sites to one which is truly fluid, but this passes through a volume with an increasing level of dynamic disorder and associated partial occupancy. Modelling of this region poses major problems, as described above, and the definition of disordered sites with low occupancy remains difficult even at atomic resolution. At which stage the occupancy and associated ADP can be defined with confidence is not yet an objective decision. In addition, refinement and modelling at this level of detail is very time consuming in terms of human intervention.

18.4.5.7. *Metal ions and other ligands in the solvent*

In general, proteins are crystallized from aqueous solutions which contain various additives, such as anions or cations (especially metals), organic solvents, including those used as cryoprotectants, and other ligands. Some of these may bind in specific or indeed non-specific sites in the ordered solvent shell, in addition to any functional binding sites of the protein. To identify such entities at limited resolution is often impossible, as the range of expected ADPs is large and there is very poor discrimination in the appearance of such sites and of water in the electron density. Atomic resolution assists in resolving ambiguities, as the interatomic distances, ADPs and occupancies are all better defined.

For metal ions, two additional criteria can be invoked. Firstly, the coordination geometry, with well defined bond lengths and angles, provides an indication of the identity of the ion, as different metals have different preferred ligand environments (Harding, 1999, 2006). The bond-valence approach is also applicable (Müller *et al.*, 2003). In addition, the value of the refined ADP and/or occupancy is helpful. Secondly, the anomalous signal in the data should reveal the presence of metal and some other non-water sites in the solvent through computation of the anomalous difference synthesis (Dauter & Dauter, 1999). This emphasizes the need to retain the anomalous signal during the collection and reduction of native data. While these approaches can be applied at lower resolution, they both become much more powerful at atomic resolution.

The presence of bound organic ligands has become especially relevant since the advent of cryogenic freezing. Compounds such as ethylene glycol and glycerol possess a number of functional hydrogen-bonding groups that can attach to sites on the protein in a defined way. Indeed, these may often bind in the active sites of enzymes such as glycosyl hydrolases, where they mimic the hydroxyl groups of the sugar substrate. It is most important to identify such moieties properly, particularly if substrate studies are to be planned successfully.

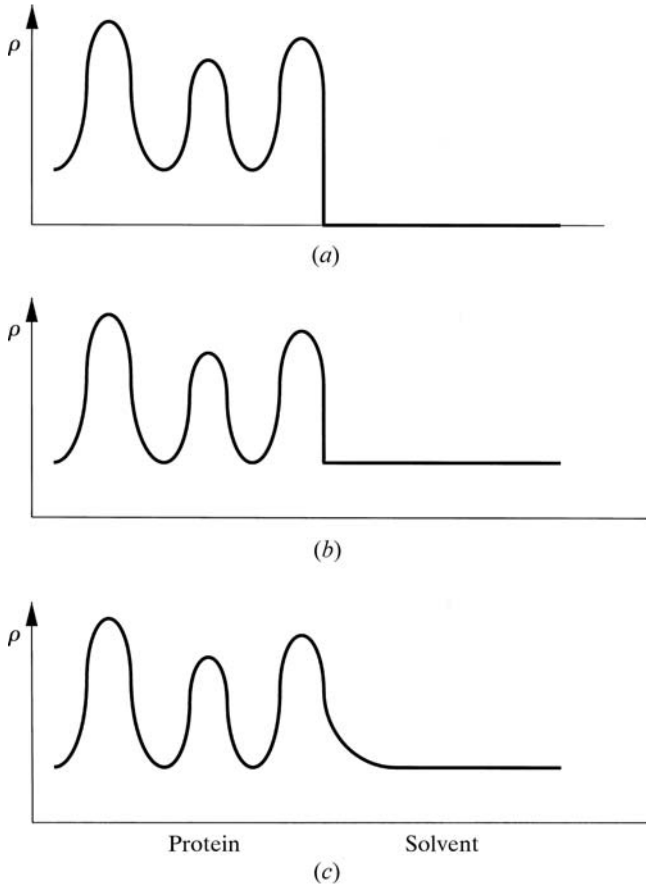


Figure 18.4.5.2
Schematic representation of the bulk-solvent models described in the text. (*a*) No bulk-solvent correction, *i.e.* solvent density set to zero. (*b*) Constant level of solvent outside the macromolecule and ordered water envelope. Here, sharp edge effects remain. (*c*) The model as in (*b*), but smoothed at the edge of a macromolecule, equivalent to the application of a *B* value to the solvent model.

18.4.5.8. *Deformation density*

X-ray structures are generally modelled using the spherical-atom approximation for the scattering, which ignores the deviation from sphericity of the outer bonding and lone-pair electrons. Extensive studies over a long period have confirmed that the so-called deformation density, representing deviation from this spherical model, can be determined experimentally using data to very high resolution, usually from 0.8 to 0.5 Å. An excellent review of this field has been provided by Coppens (1997). The observed deviations can be compared with those expected from the available theories of chemical bonding and the densities derived therefrom.

The application of atomic resolution analysis to proteins allowed the observation of the deformation density in macromolecules (Lamzin *et al.*, 1999). Data for two proteins were analysed: crambin (molecular weight 6 kDa) at 0.67 Å resolution and a subtilisin (molecular weight 30 kDa) at 0.9 Å resolution. Significant and interpretable deformation density could not be observed for the individual residues. However, on averaging the density over 40 peptide units for crambin and more than 250 for the subtilisin, the deformation density within the peptide unit was clearly visible and could be related to the expected bonding features in these units. This shows the real power of atomic resolution crystallography, which can reveal features containing no more than 0.2 e Å^{-3}.

Deformation density studies are now being applied to many polypeptides (Jelsch *et al.*, 1998; Koritsanszky *et al.*, 2002; Pichon-

Pesme *et al.*, 2004; Afonine *et al.*, 2007; Zarychta *et al.*, 2007). It has been observed that details of deformation densities are most clearly revealed when only the highest resolution (so-called high order) terms are included in the refinement and the Fourier maps (Coppens, 1997). This is reported to result from the deconvolution of the effects of the anisotropic ADPs, which can to some extent take up the fine features corresponding to the deformation density (bonding electrons and lone pairs). After proper modelling of the deformation density features, the overall model is refined against all the data.

18.4.6. Quality assessment of the model

The refinement of proteins at resolutions lower than atomic depends upon the use of restraints on the geometry and ADPs. The almost exclusively used library of target geometric restraints for refinement and validation of protein structures (Chapter 18.3) is derived from structures of amino acids and peptides in the Cambridge Crystallographic Data Centre's small-molecule crystal structure database (Allen *et al.*, 1979). Stereochemical parameters, such as conformational angles φ, ψ, should ideally not be restrained, as they allow independent validation of the model. As stated in Section 18.4.4.2, these restraints are required even at atomic resolution to maintain the chemical integrity of flexible regions, although their impact will be limited on ordered regions.

Owing to the excess of accurate X-ray observations over parameters at atomic resolution, extensive validation of individual structures should be less challenging than for those at lower resolution. It is hard to achieve an R factor around 10% with an incorrect model. However, considerable attention needs to be given to detail and great care taken to avoid over-interpretation, especially of the flexible regions.

An analysis of eight structures determined at atomic resolution some years ago (EU 3-D Validation Network, 1998) indicated that they follow the expected rules of chemistry more closely than those of lower-resolution analyses in the PDB, confirming that atomic resolution indeed provides more precise coordinates. A subsequent analysis of ten structures at ultra-high resolution, 0.8 Å or better (Jaskolski *et al.*, 2007) confirmed these conclusions but identified a few possible adjustments to some targets. Following this analysis, Karplus *et al.* (2008) proposed that protein stereochemistry is context dependent, *e.g.* it differs in detail between α-helices and β-strands, and that this should be reflected in future target libraries. Thus, the availability of atomic resolution structures will provide a more objective basis for the construction of such libraries.

18.4.7. Relation to biological chemistry

A question arises as to what biological issues are addressed by analysis of macromolecular structures at atomic resolution. For any protein, the overall structure of its fold, and hence its homology with other proteins, can already be provided by analyses at low to medium resolution. However, proteins are the active entities of cells and carry out recognition of other macromolecules, ligand binding and catalytic roles that depend upon subtle details of chemistry, for which accurate positioning of the atoms is required. Even at atomic resolution, the accuracy of structural definition is less than what would ideally be required for the changes observed during a chemical reaction. At lower resolutions, structure–function relations require yet further extrapolation of the experimental data.

To understand the function of many macromolecules, such as enzymes, it is not sufficient to determine the structure of a single state. Alongside the native structure, those of various complexes will also be required. The differences between the states provide additional information on the functionality. For an understanding of the chemistry involved, atomic resolution has tremendous advantages in terms of accuracy, as reliable judgments can be based on the experimental data alone.

Advantages of atomic resolution include the following:

(1) The positions of all atoms that possess defined conformations are more accurately defined. This means that all bond lengths and angles in the structure have lower standard uncertainties (EU 3-D Validation Network, 1998). For regions of the molecule where the conformation is representative of the norm, this is of purely quantitative significance, but where the stereochemistry deviates from the expected value this accuracy takes on a special significance, which poses questions to the theoretical chemist. Such deviations from standard geometry often play an important role in biological function.

(2) The better the ADP definition, notably its anisotropy, the greater the insight into the static or thermal flexibility of individual regions of the molecule. Macromolecules are crucially dependent upon flexibility for properties such as induced fit in substrate or ligand recognition, allosteric responses or responses to the biological environment. More detailed definition of the position and mobility of flexible regions may be assisted by atomic resolution analysis.

(3) A few amino-acid side chains play an active role in catalysis. Those that do include histidine, aspartic and glutamic acids and serine, all through protonation–deprotonation events, and hydrogen atoms are crucial to their function. Hydrogen atoms are usually treated as riding on their parent atoms and should be included in the model, even at medium resolution. Unfortunately, those hydrogen atoms that are of interest can only rarely be treated as rigidly bonded at a predictable position. However, atomic resolution allows many hydrogen atoms to be clearly identified in the refined electron density. In addition, the presence or absence of hydrogen may be inferred by accurate estimation of the bond lengths between atoms, *e.g.* within the carboxylate groups.

(4) The relative orientation of reacting moieties is crucial to enzyme catalysis. If chemical hypotheses of mechanism are to be subjected to appropriate Popperian scrutiny (Popper, 1959), then precise definition of atomic coordinates in native and complex structures is necessary.

(5) Enzyme catalysis provides a reduction of the activation energy of the reaction, which can be achieved by distortion of the conformation of the substrate bound to the enzyme (the so-called Michaelis complex) towards the transition state or by the stabilization of the latter by the enzyme. For both, the study of complexes of inhibitors or substrate analogues at a sufficient resolution to clarify the fine detail of the structures is required.

(6) Adaptation of the enzyme to the substrate is postulated by the induced-fit theory of catalysis. The level of adjustment can be very small, and energy calculations again require that this be precisely defined.

(7) In metalloproteins, the ligand field, and hence geometry and bond lengths, around the metal ion are essential indicators of any variation in valence electrons between different states. For example, bond lengths between oxidized and reduced states of metal ions vary by the order of 0.1 Å or less, and

clear distinction between alternative oxidation states requires an accuracy only provided by atomic resolution.

Almost all atomic resolution analyses require data recorded from cryogenically frozen crystals. This does pose some problems of biological relevance, as proteins *in vivo* have adapted to operate at ambient cellular temperatures. The structure that is required is that of the protein and surrounding solvent at the corresponding temperature. The trade-off is that cryogenic structures may be better defined, but that they are only so because of the increased order of the protein and solvent at low temperature. This has to be weighed against the lack of fine detail in a medium-resolution analysis at room temperature.

A question often raised with regard to the worth of atomic resolution data concerns the effort required in refining a protein at such resolution. To define all details, such as alternative conformations, hydrogen-atom positions and solvent, is certainly time consuming, especially if an anisotropic model is adopted. However, the advantages outweigh the disadvantages, as even if a full anisotropic model is not refined to exhaustion, nevertheless all density maps will be clearer if the resolution is better, resulting in an improved definition of the features of interest.

18.4.8. Practical strategies

Pioneering work was carried out by Teeter and colleagues on crambin, using data recorded on this small and highly stable protein using a conventional diffractometer (Teeter *et al.*, 1993). Some of the earliest atomic resolution structures using data from a synchrotron source with an imaging plate detector included rubredoxin at 1.0 Å (Dauter *et al.*, 1992), ribonuclease Sa at 1.1 Å (Sevcik *et al.*, 1996) and triclinic lysozyme at 0.9 Å resolution (Walsh *et al.*, 1998), at room temperature for the rubredoxin and ribonuclease, and 100 K for the lysozyme. The strategy used involved the application of conventional restrained least-squares or maximum-likelihood techniques in the early stages of refinement, followed by switching over to *SHELXL* to introduce a full anisotropic model and riding hydrogen atoms.

Subsequent structures have almost exclusively used cryogenically vitrified crystals to minimize the effects of radiation damage on high-intensity SR beamlines. A review of the field (Dauter, 2003) reported over 100 structures at better than 1.2 Å spacing, and in January 2011 there were around 1450 such coordinate sets in the PDB.

Initially, only *SHELXL* was able to refine macromolecular structures with anisotropic ADPs, riding hydrogen atoms and appropriate stereochemical restraints. Subsequently, such options have been implemented in programs such as *REFMAC* and *PHENIX.REFINE*. The use of fast Fourier algorithms gives these programs some speed advantage over *SHELXL*, but this is now of less importance given the power of contemporary computers. The key advantage of *SHELXL* is the ability to perform full- (or at least block-) matrix inversion, and hence extract error estimates for each individual parameter.

It is not straightforward to provide hard and fast rules for the optimum strategy. The power of refinement at atomic resolution lies in the sheer number of X-ray observations which drive the minimization to a global minimum. The dangers of bias in the final model and of false minima can largely be ignored. As a guideline the following steps would seem to be a common-sense approach when data to 1.2 Å or better are available.

(1) Refine an initial isotropic model at modest resolution with one of the conventional programs to ensure the correct global minimum has been reached. This step was probably more appropriate in the days when computing resources were more limiting, and could probably be omitted today.

(2) Introduce an anisotropic model with riding hydrogen atoms using one of the maximum-likelihood programs such as *REFMAC*, *PHENIX.REFINE* or *SHELXL* (with the conjugate-gradient option), and refine against the experimental data, omitting around 1000–2000 reflections for validation by $R_{\rm free}$. It is *not* necessary to omit 5% of the data, particularly at this resolution.

(3) Complete the water model using one of the autobuild options within or associated with (*e.g. ARP/wARP* or *COOT*) the refinement program. This is an area amenable to considerable automation in the future to allow partially occupied sites and overlapping networks to be defined without user intervention.

(4) Introduce alternative conformations for the protein. Much of this is also amenable to automation, especially for the side chains, where programs such as *COOT* can already suggest likely alternate rotamers.

(5) Inspect appropriately weighted electron-density maps for features to be added, deleted or moved. This includes ligands, metal atoms, ions, cryprotectants and other additives. Current developments in programs such as *COOT* are increasingly automating this step. Software libraries that provide accurate descriptions of the ligand chemistry and geometry are essential, and are the subject of intense development. It is vital that structures are deposited in the PDB with the correct geometrical restraints. To some extent, steps (3)–(5) go in parallel, and indeed their order is not vital.

(6) Consider the *SHELXL* full-matrix option for the final refinement cycles. This will provide estimates of the accuracy of the individual parameters, not available from other programs.

(7) As a final step, run some cycles with all reflections including those previously used for $R_{\rm free}$, but without changing the model parameters. At all stages, it will remain necessary to impose restraints to avoid poor behaviour of any flexible regions. As the resolution goes beyond 1 Å, the effect of such restraints on good parts of the structure becomes minimal.

In summary, at atomic resolution models require special attention to detail, reflecting the wealth of information with regard to features such as alternative conformations of both side and main chains, and extensive ordered water structure often with partial occupancy and overlapping networks. With the present software, this can be a lengthy operation, and indeed this can persuade some scientists to work at lower, less informative, resolution to speed up and simplify the analysis. Many of the presently tedious manual tasks in terms of model construction are, however, amenable to extensive automation. As the number of atomic resolution structures increases, we can expect the refinement and modelling process to be greatly simplified with new software algorithms.

References

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., Lunin, V. Y. & Urzhumtsev, A. (2007). *On macromolecular refinement at subatomic resolution with interatomic scatterers*. *Acta Cryst.* D**63**, 1194–1197.

Agarwal, R. C. (1978). *A new least-squares refinement technique based on the fast Fourier transform algorithm*. *Acta Cryst.* A**34**, 791–809.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Andersson, K. M. & Hovmöller, S. (1998). *The average atomic volume and density of proteins*. *Z. Kristallogr.* **213**, 369–373.

Blessing, R. H. (1997). *LOCSCL: a program to statistically optimize local scaling of single-isomorphous-replacement and single-wavelength-anomalous-scattering data*. *J. Appl. Cryst.* **30**, 176–177.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, California, London: Addison-Wesley.

Bricogne, G. & Irwin, J. J. (1996). *Maximum-likelihood structure refinement: theory and implementation within BUSTER+TNT*. In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992*a*). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR manual*. Version 3.1. New Haven: Yale University.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR system: a new software suite for macromolecular structure determination*. *Acta Cryst.* D**54**, 905–921.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography*. *Acta Cryst.* D**50**, 760–763.

Coppens, P. (1997). *X-ray Charge Densities and Chemical Bonding*. International Union of Crystallography and Oxford University Press.

Cowtan, K. D. & Main, P. (1998). *Miscellaneous algorithms for density modification*. *Acta Cryst.* D**53**, 487–493.

Cruickshank, D. W. J. (1999*a*). *Remarks about protein structure precision*. *Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. (1999*b*). *Remarks about protein structure precision. Erratum*. *Acta Cryst.* D**55**, 1108.

Dauter, Z. (2003). *Protein structures at atomic resolution*. *Methods Enzymol.* **368**, 288–337.

Dauter, Z. & Dauter, M. (1999). *Anomalous signal of solvent bromides used for phasing of lysozyme*. *J. Mol. Biol.* **289**, 93–101.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution*. *Curr. Opin. Struct. Biol.* **7**, 681–688.

Dauter, Z., Sieker, L. C. & Wilson, K. S. (1992). *Refinement of rubredoxin from Desulfovibrio vulgaris at 1.0 Å with and without restraints*. *Acta Cryst.* B**48**, 42–59.

Dauter, Z., Wilson, K. S., Sieker, L. C., Meyer, J. & Moulis, J.-M. (1997). *Atomic resolution (0.94 Å) structure of Clostridium acidurici ferredoxin. Detailed geometry of [4Fe-4S] clusters in a protein*. *Biochemistry*, **36**, 16065–16073.

Diamond, R. (1971). *A real-space refinement procedure for proteins*. *Acta Cryst.* A**27**, 436–452.

Driessen, H., Haneef, M. I. J., Harris, G. W., Howlin, B., Khan, G. & Moss, D. S. (1989). *RESTRAIN: restrained structure-factor least-squares refinement program for macromolecular structures*. *J. Appl. Cryst.* **22**, 510–516.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics*. *Acta Cryst.* D**60**, 2126–2132.

EU 3-D Validation Network (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures*. *J. Mol. Biol.* **276**, 417–436.

French, S. & Wilson, K. S. (1978). *On the treatment of negative intensity observations*. *Acta Cryst.* A**34**, 517–525.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents*. *J. Am. Chem. Soc.* **118**, 519–528.

Harding, M. M. (1999). *The geometry of metal–ligand interactions relevant to proteins*. *Acta Cryst.* D**55**, 1432–1443.

Harding, M. M. (2006). *Small revisions to predicted distances around metal sites in proteins*. *Acta Cryst.* D**62**, 678–682.

Herzberg, O. & Sussman, J. L. (1983). *Protein model building by the use of a constrained–restrained least-squares procedure*. *J. Appl. Cryst.* **16**, 144–150.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them?* *Acta Cryst.* D**63**, 611–620.

Jelsch, C., Pichon-Pesme, V., Lecomte, C. & Aubry, A. (1998). *Transferability of multipole charge-density parameters: application to very high resolution oligopeptide and protein structures*. *Acta Cryst.* D**54**, 1306–1318.

Jiang, J. S. & Brunger, A. T. (1994). *Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures*. *J. Mol. Biol.* **243**, 100–115.

Johnson, C. K. (1976). *ORTEPII. A FORTRAN thermal-ellipsoid plot program for crystal structure illustration*. Report ORNL-5138. Oak Ridge National Laboratory, Tennessee, USA.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models*. *Acta Cryst.* A**47**, 110–119.

Karplus, P. A., Shapovalov, M. V., Dunbrack, R. L. Jr & Berkholz, D. S. (2008). *A forward-looking suggestion for resolving the stereochemical restraints debate: ideal geometry functions*. *Acta Cryst.* D**64**, 335–336.

Koritsanszky, T., Volkov, A. & Coppens, P. (2002). *Aspherical-atom scattering factors from molecular wave functions. 1. Transferability and conformation dependence of atomic electron densities of peptides within the multipole formalism*. *Acta Cryst.* A**58**, 464–472.

Konnert, J. H. & Hendrickson, W. A. (1980). *A restrained-parameter thermal-factor refinement procedure*. *Acta Cryst.* A**36**, 344–350.

Lamzin, V. S., Morris, R. J., Dauter, Z., Wilson, K. S. & Teeter, M. M. (1999). *Experimental observation of bonding electrons in proteins*. *J. Biol. Chem.* **274**, 20753–20755.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography*. *Methods Enzymol.* **277**, 269–305.

Lebedev, A. A., Vagin, A. A. & Mushudov, G. N. (2006). *Intensity statistics in twinned crystals with examples from the PDB*. *Acta Cryst.* D**62**, 83–95.

Matthews, B. W. (1968). *Solvent content in protein crystals*. *J. Mol. Biol.* **33**, 491–497.

Morris, R. J., Blanc, E. & Bricogne, G. (2004). *On the interpretation and use of* $\langle|E|^2\rangle(d^*)$ *profiles*. *Acta Cryst.* D**60**, 227–240.

Morris, R. J. & Bricogne, G. (2003). *Sheldrick's 1.2 Å rule and beyond*. *Acta Cryst.* D**59**, 615–617.

Müller, P., Köpke, S. & Sheldrick, G. M. (2003). *Is the bond-valence method able to identify metal atoms in protein structures?* *Acta Cryst.* D**59**, 32–37.

Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *The effect of overall anisotropic scaling in macromolecular refinement*. In *CCP4 Newsletter on Protein Crystallography*, **35**, 37–42.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method*. *Acta Cryst.* D**53**, 240–255.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Efficient anisotropic refinement of macromolecular structures using FFT*. *Acta Cryst.* D**55**, 247–255.

O'Hagan, A. (1994). *Kendal's Advanced Theory of Statistics; Bayesian Inference*, Vol. 2B. Cambridge: Arnold/Hodder Headline/Cambridge University Press.

Painter, J. & Merritt, E. A. (2006). *Optimal description of a protein structure in terms of multiple groups undergoing TLS motion*. *Acta Cryst.* D**62**, 439–450.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. *Acta Cryst.* A**52**, 659–668.

Pichon-Pesme, V., Jelsch, C., Guillot, B. & Lecomte, C. (2004). *A comparison between experimental and theoretical aspherical-atom scattering factors for charge-density refinement of large molecules*. *Acta Cryst.* A**60**, 204–208.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Prince, E. & Boggs, P. T. (2004). *International Tables for Crystallography*, Vol. C, ch. 8.1. Dordrecht: Kluwer Academic Publishers.

Schomaker, V. & Trueblood, K. N. (1968). *On the rigid-body motion of molecules in crystals*. *Acta Cryst.* B**24**, 63–76.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Statistical descriptors in crystallography. II. Report of a working group on expression of uncertainty in measurement*. *Acta Cryst.* A**51**, 565–569.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Ribonuclease from Streptomyces aureofaciens at atomic resolution. Acta Cryst.* D**52**, 327–344.

Sheldrick, G. M. (1990). *Phase annealing in SHELX-90: direct methods for larger structures. Acta Cryst.* A**46**, 467–473.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high-resolution refinement. Methods Enzymol.* **277**, 319–343.

Sheriff, S. & Hendrickson, W. A. (1987). *Description of overall anisotropy in diffraction from macromolecular crystals. Acta Cryst.* A**43**, 118–121.

Steiner, R. A., Lebedev, A. A. & Murshudov, G. N. (2003). *Fisher's information in maximum-likelihood macromolecular crystallographic refinement. Acta Cryst.* D**59**, 2114–2124.

Strokopytov, B. V. (2008). *How to multiply a matrix of normal equations by an arbitrary vector using FFT. Acta Cryst.* A**64**, 601–612.

Stuart, A., Ord, K. J. & Arnold, S. (1999). *Kendall's Advanced Theory of Statistics; Classical Inference and Linear Model*, Vol. 2A. London, Sydney, Auckland: Arnold/Hodder Headline.

Teeter, M. M., Roe, S. M. & Heo, N. H. (1993). *Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 K. J. Mol. Biol.* **230**, 292–311.

Ten Eyck, L. F. (1973). *Crystallographic fast Fourier transforms. Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* A**33**, 486–492.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard*. *Acta Cryst*. D**64**, 61–69.

Tickle, I. J. (2007). *Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values*. *Acta Cryst.* D**63**, 1274–1281.

Tronrud, D. E. (1997). *TNT refinement package. Methods Enzymol.* **277**, 243–268.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst*. D**55**, 191–205.

Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Refinement of triclinic hen egg-white lysozyme at atomic resolution. Acta Cryst.* D**54**, 522–546.

Wilson, A. J. C. (1942). *Determination of absolute from relative X-ray data intensities. Nature (London)*, **150**, 151–152.

Winn, M. D., Isupov, M. N. & Murshudov, G. N. (2001). *Use of TLS parameters to model anisotropic displacements in macromolecular refinement*. *Acta Cryst.* D**57**, 122–133.

Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. (2008). *Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures*. *FEBS J.* **275**, 1–21.

Yeates, T. O. (1997). *Detecting and overcoming crystal twinning*. *Methods Enzymol.* **276**, 344–358.

Zarychta, B., Pichon-Pesme, V., Guillot, B., Lecomte, C. & Jelsch, C. (2007). *On the application of an experimental multipolar pseudo-atom library for accurate refinement of small-molecule and protein crystal structures*. *Acta Cryst.* A**63**, 108–125.

International Tables for Crystallography (2012). Vol. F, Chapter 18.5, pp. 499–511.

Chapter 18.5. Coordinate uncertainty

D. W. J. Cruickshank

18.5.1. Introduction

18.5.1.1. Background

Even in 1967 when the first few protein structures had been solved, it would have been hard to imagine a time when the best protein structures would be determined with a precision approaching that of small molecules. That time was reached during the 1990s. Consequently, the methods for the assessment of the precision of small molecules can be extended to good-quality protein structures.

The key idea is simply stated. At the conclusion and full convergence of a least-squares or equivalent refinement, *the estimated variances and covariances of the parameters may be obtained through the inversion of the least-squares full matrix.*

The inversion of the full matrix for a large protein is a gigantic computational task, but it is being accomplished in a rising number of cases. Alternatively, approximations may be sought. Often these can be no more than rough order-of-magnitude estimates. Some of these approximations are considered below.

Caveat. Quite apart from their large numbers of atoms, protein structures show features differing from those of well ordered small-molecule structures. Protein crystals contain large amounts of solvent, much of it not well ordered. Parts of the protein chain may be floppy or disordered. All natural protein crystals are noncentrosymmetric, hence the simplifications of error assessment for centrosymmetric structures are inapplicable. The effects of incomplete modelling of disorder on phase angles, and thus on parameter errors, are not addressed explicitly in the following analysis. Nor does this analysis address the quite different problem of possible gross errors or misplacements in a structure, other than by their indication through high B values or high coordinate standard uncertainties. These various difficulties are, of course, reflected in the values of $\Delta|F|$ used in the precision estimates.

On the problems of structure validation see Part 21 of this volume and Dodson (1998).

Some structure determinations do make a first-order correction for the effects of disordered solvent on phase angles by application of Babinet's principle of complementarity (Langridge *et al.*, 1960; Moews & Kretsinger, 1975; Tronrud, 1997). Babinet's principle follows from the fact that if $\rho(\mathbf{x})$ is constant throughout the cell, then $F(\mathbf{h}) = 0$, except for $F(\mathbf{0})$. Consequently, if the cell is divided into two regions C and D, $F_C(\mathbf{h}) = -F_D(\mathbf{h})$. Thus if D is a region of disordered solvent, $F_D(\mathbf{h})$ can be estimated from $-F_C(\mathbf{h})$. A first approximation to a disordered model may be obtained by placing negative point-atoms with very high Debye B values at all the ordered sites in region C. This procedure provides some correction for very low resolution planes. Alternatively, corrections are sometimes made by a mask bulk solvent model (Jiang & Brünger, 1994).

The application of restraints in protein refinement does not affect the key idea about the method of error estimation. A simple model for restrained refinement is analysed in Section 18.5.3, and the effect of restraints is discussed in Section 18.5.4 and later.

Much of the material in this chapter is drawn from a Topical Review published in *Acta Crystallographica*, Section D (Cruickshank, 1999).

Protein structures exhibiting noncrystallographic symmetry are not considered in this chapter.

18.5.1.2. Accuracy and precision

A distinction should be made between the terms *accuracy* and *precision.* A single measurement of the magnitude of a quantity differs by error from its unknown true value λ. In statistical theory (Cruickshank, 1959), the fundamental supposition made about errors is that, for a given experimental procedure, the possible results of an experiment define the probability density function $f(x)$ of a *random variable.* Both the true value λ and the probability density $f(x)$ are unknown. The problem of assessing the accuracy of a measurement is thus the double problem of estimating $f(x)$ and of assuming a relation between $f(x)$ and λ.

Precision relates to the function $f(x)$ and its spread.

The problem of what relationship to assume between $f(x)$ and the true value λ is more subtle, involving particularly the question of *systematic errors.* The usual procedure, after correcting for known systematic errors, is to suppose that some typical property of $f(x)$, often the mean, is the value of λ. No repetition of the same experiment will ever reveal the systematic errors, so statistical estimates of precision take into account only random errors. Empirically, systematic errors can be detected only by remeasuring the quantity with a different technique.

Care is needed in reading older papers. The word accuracy was sometimes intended to cover both random and systematic errors, or it may cover only random errors in the above sense of precision (known systematic errors having been corrected).

In recent years, the well established term *estimated standard deviation* (e.s.d.) has been replaced by the term *standard uncertainty* (s.u.). (See Section 18.5.2.3 on statistical descriptors.)

18.5.1.3. Effect of atomic displacement parameters (or 'temperature factors')

It is useful to begin with a reminder that the Debye $B = 8\pi^2\langle u^2\rangle$, where u is the atomic displacement parameter. If $B = 80$ Å^2, the r.m.s. amplitude is 1.01 Å. The centroid of an atom with such a B is unlikely to be precisely determined. For $B = 40$ Å^2, the 0.71 Å r.m.s. amplitude of an atom is approximately half a C—N bond length. For $B = 20$ Å^2, the amplitude is 0.50 Å. Even for $B = 5$ Å^2, the amplitude is 0.25 Å. The size of the atomic displacement amplitudes should always be borne in mind when considering the precision of the position of the centroid of an atom.

Scattering power depends on $\exp[-2B(\sin\theta/\lambda)^2] = \exp[-B/(2d^2)]$. For $B = 20$ Å^2 and $d = 4$, 2 or 1 Å, this factor is 0.54, 0.08 or 0.0001. For $d = 2$ Å and $B = 5$, 20 or 80 Å^2, the factor is again 0.54, 0.08 or 0.0001. The scattering power of an atom thus depends very strongly on B and on the resolution $d = 1/s = \lambda/2\sin\theta$. Scattering at high resolution (low d) is dominated by atoms with low B.

Copyright © 2012 International Union of Crystallography

An immediate consequence of the strong dependence of scattering power on B is that the standard uncertainties of atomic coordinates also depend very strongly on B, especially between atoms of different B within the same structure.

[An IUCr Subcommittee on Atomic Displacement Parameter Nomenclature (Trueblood *et al.*, 1996) has recommended that the phrase 'temperature factor', though widely used in the past, should be avoided on account of several ambiguities in its meaning and usage. The Subcommittee also discourages the use of B and the anisotropic tensor $\mathbf{B}$ in favour of $\langle u^2 \rangle$ and $\mathbf{U}$, on the grounds that the latter have a more direct physical significance. The present author concurs (Cruickshank, 1956, 1965). However, as the use of B or $B_{\rm eq}$ is currently so widespread in biomolecular crystallography, this chapter has been written in terms of B.]

18.5.2. The least-squares method

18.5.2.1. *The normal equations*

In the unrestrained least-squares method, the residual

$$R = \sum_3 w(hkl)\Delta^2(hkl) \tag{18.5.2.1}$$

is minimized, where Δ is either $|F_o| - |F_c|$ for R_1 or $|F_o|^2 - |F_c|^2$ for R_2, and $w(hkl)$ is chosen appropriately. The summation is over crystallographically independent planes.

When R is a minimum with respect to the parameter u_j, $\partial R/\partial u_j = 0$, *i.e.*,

$$\sum_3 w\Delta(\partial\Delta/\partial u_j) = 0. \tag{18.5.2.2}$$

For R_1, $\partial\Delta/\partial u_j = -\partial|F_c|/\partial u_j$; for R_2, $\partial\Delta/\partial u_j = -2|F_c|\partial|F_c|/\partial u_j$. The n parameters have to be varied until the n conditions (18.5.2.2) are satisfied. For a trial set of the u_j close to the correct values, we may expand Δ as a function of the parameters by a Taylor series to the first order. Thus for R_1,

$$\Delta(\mathbf{u} + \mathbf{e}) = \Delta(\mathbf{u}) - \sum_i \varepsilon_i(\partial|F_c|/\partial u_i), \tag{18.5.2.3}$$

where ε_i is a small change in the parameter u_i, and $\mathbf{u}$ and $\mathbf{e}$ represent the whole sets of parameters and changes. The minus sign occurs before the summation, since $\Delta = |F_o| - |F_c|$, and the changes in $|F_c|$ are being considered.

Substituting (18.5.2.3) in (18.5.2.2), we get the *normal equations* for R_1,

$$\begin{aligned}\sum_i \varepsilon_i \left[\sum_3 w(\partial|F_c|/\partial u_i)(\partial|F_c|/\partial u_j)\right]&\\ = \sum_3 w\Delta(\partial|F_c|/\partial u_j).&\end{aligned} \tag{18.5.2.4}$$

There are n of these equations for $j = 1, \ldots, n$ to determine the n unknown ε_j.

For R_2 the normal equations are

$$\begin{aligned}\sum_i \varepsilon_i \left[\sum_3 w(\partial|F_c|^2/\partial u_i)(\partial|F_c|^2/\partial u_j)\right]&\\ = \sum_3 w\Delta(\partial|F_c|^2/\partial u_j).&\end{aligned} \tag{18.5.2.5}$$

Both forms of the normal equations can be abbreviated to

$$\sum_i \varepsilon_i a_{ij} = b_j. \tag{18.5.2.6}$$

For the values of $\partial|F_c|/\partial u_j$ for common parameters see, *e.g.*, Cruickshank (1970).

Some important points in the derivation of the standard uncertainties of the refined parameters can be most easily understood if we suppose that the matrix a_{ij} can be approximated by its diagonal elements. Each parameter is then determined by a single equation of the form

$$\varepsilon_i \sum_3 wg^2 = \sum_3 wg\Delta, \tag{18.5.2.7}$$

where $g = \partial|F_c|/\partial u_i$ or $\partial|F_c|^2/\partial u_i$. Hence

$$\varepsilon_i = \left(\sum_3 wg\Delta\right) \bigg/ \left(\sum_3 wg^2\right). \tag{18.5.2.8}$$

At the conclusion of the refinement, when R is a minimum, the variance (square of the s.u.) of the parameter u_i due to uncertainties in the Δ's is

$$\sigma_i^2 = \left[\sum_3 w^2 g^2 \sigma^2(F)\right] \bigg/ \left(\sum_3 wg^2\right)^2. \tag{18.5.2.9}$$

If the weights have been chosen as $w(hkl) = 1/\sigma^2(|F_{hkl}|)$ or $1/\sigma^2(|F_{hkl}|^2)$, this simplifies to

$$\sigma_i^2 = 1 \bigg/ \left(\sum_3 wg^2\right) = 1/a_{ii}, \tag{18.5.2.10}$$

which is appropriate for absolute weights. Equation (18.5.2.10) provides an s.u. for a parameter relative to the s.u.'s $\sigma(|F|)$ or $\sigma(|F|^2)$ of the observations.

In general, with the full matrix a_{ij} in the normal equations,

$$\sigma_i^2 = (a^{-1})_{ii}, \tag{18.5.2.11}$$

where $(a^{-1})_{ii}$ is an element of the matrix inverse to a_{ij}. The covariance of the parameters u_i and u_j is $\operatorname{cov}(i, j) \equiv \sigma_i\sigma_j\operatorname{correl}(i, j) = (a^{-1})_{ij}$.

18.5.2.2. *Weights*

In the early stages of refinement, artificial weights may be chosen to accelerate refinement. In the final stages, the weights must be related to the precision of the structure factors if parameter variances are being sought. There are two distinct ways, covering two ranges of error, in which this may be done.

(1) The weights for R_1, say, may reflect the precision of the $|F_o|$, so that $w(hkl) = 1/\sigma^2(|F_{hkl}|)$, where σ^2 is the estimated variance of $|F_o|$ due to a specific class of experimental uncertainties. These absolute weights are derived from an analysis of the experiment. Weights chosen in this way lead to estimated parameter variances $\sigma_i^2 = (a^{-1})_{ii}$, (18.5.2.11), which cover only the specific class of experimental uncertainties.

(2) The weights may reflect the trends in the $|\Delta| \equiv \||F_o| - |F_c|\|$. A weighting function with a small number of parameters is chosen so that the averages of $w\Delta^2$ are constant when the set of $w\Delta^2$ values is analysed in any pertinent fashion (*e.g.* in bins of increasing $|F_o|$ and $2\sin\theta/\lambda$). Weights chosen in this way are relative weights, and the expression for the parameter variances needs a scaling factor,

$$S^2 = \left(\sum_3 w\Delta^2\right) \bigg/ (n_{\rm obs} - n_{\rm params}). \tag{18.5.2.12}$$

Hence, in the full-matrix case,

$$\sigma_i^2 = \left[\left(\sum_3 w\Delta^2\right) \bigg/ (n_{\rm obs} - n_{\rm params})\right](a^{-1})_{ii}, \tag{18.5.2.13}$$

which allows for all random experimental errors, such systematic experimental errors as cannot be simulated in the $|F_c|$ and imperfections in the calculated model.

18.5.2.3. Statistical descriptors and goodness of fit

In recent years, there have been developments and changes in statistical nomenclature and usage. Many aspects are summarised in the reports of the IUCr Subcommittee on Statistical Descriptors in Crystallography (Schwarzenbach *et al.*, 1989, 1995). In the second report, *inter alia*, the Subcommittee emphasizes the terms *uncertainty* and *standard uncertainty* (s.u.). The latter is a replacement for the older term *estimated standard deviation* (e.s.d.). The Subcommittee classify uncertainty components in two categories, based on their method of evaluation: type A, estimated by the statistical analysis of a series of observations, and type B, estimated otherwise. As an example of the latter, a type B component could allow for doubts concerning the estimated shape and dimensions of the diffracting crystal and the subsequent corrections made for absorption.

The square root S of the expression S^2, (18.5.2.12) above, is called the *goodness of fit* when the weights are the reciprocals of the absolute variances of the observations.

One recommendation in the second report does call for comment here. While agreeing that formulae like (18.5.2.13) lead to conservative estimates of parameter variances, the report suggests that this practice is based on the questionable assumption that the variances of the observations by which the weights are assigned are relatively correct but uniformly underestimated. When the goodness of fit $S > 1$, then either the weights or the model or both are suspect.

Comment is needed. The account in Section 18.5.2.2 describes two distinct ways of estimating parameter variances, covering two ranges of error. The kind of weights envisaged in the reports (based on variances of type A and/or of type B) are of a class described for method (1). They are not the weights to be used in method (2) (though they may be a component in such weights). Method (2) implicitly assumes from the outset that there are experimental errors, some covered and others not covered by method (1), and that there are imperfections in the calculated model (as is obviously true for proteins). Method (2) avoids exploring the relative proportions and details of these error sources and aims to provide a realistic estimate of parameter uncertainties which can be used in external comparisons. It can be formally objected that method (2) does not conform to the criteria of random-variable theory, since clearly the Δ's are partially correlated through the remaining model errors and some systematic experimental errors. But it is a useful procedure. Method (1) on its own would present an optimistic view of the reliability of the overall investigation, the degree of optimism being indicated by the inverse of the goodness of fit (18.5.2.12). In method (2), if the weights are on an arbitrary scale, then S^2 can have an arbitrary value.

For an advanced-level treatment of many aspects of the refinement of structural parameters, see Part 8 of *International Tables for Crystallography*, Volume C (2004). The detection and treatment of systematic error are discussed in Chapter 8.5 therein.

18.5.3. Restrained refinement

18.5.3.1. Residual function

Protein structures are often refined by a restrained refinement program such as *PROLSQ* (Hendrickson & Konnert, 1980). Here, a function of the type

$$R' = \sum w_h(\Delta F)^2 + \sum w_{\rm geom}(\Delta Q)^2 \qquad (18.5.3.1)$$

is minimized, where Q denotes a geometrical restraint such as a bond length. Formally, all one is doing is extending the list of observations. One is adding to the protein diffraction data geometrical data from a stereochemical dictionary such as that of Engh & Huber (1991). A chain C—N bond length may be known from the dictionary with much greater precision $1/w_{\rm geom}^{1/2}$, say 0.02 Å, than from an unrestrained diffraction-data-only protein refinement.

In a high-resolution unrestrained refinement of a small molecule, the standard uncertainty (s.u.) of a bond length $A-B$ is often well approximated by

$$\sigma(l) = (\sigma_A^2 + \sigma_B^2)^{1/2}. \qquad (18.5.3.2)$$

However, in a protein determination $\sigma(l)$ is often much smaller than either σ_A or σ_B because of the excellent information from the stereochemical dictionary, which correlates the positions of A and B.

Laying aside computational size and complexity, the protein precision problem is straightforward in principle. When a restrained refinement has converged to an acceptable structure and the shifts in successive rounds have become negligible, invert the full matrix. The inverse matrix immediately yields estimates of the variances and covariances of all parameters.

The dimensions of the matrix are the same whether or not the refinement is restrained. The full matrix will be rather sparse, but not nearly as sparse as in a small-molecule refinement. For the purposes of Section 18.5.3, it is irrelevant whether the residual for the diffraction data is based on $|F|$ or $|F|^2$. On the relative weighting of the diffraction and restraint terms, see Section 18.5.3.3.

18.5.3.2. A very simple protein model

Some aspects of restrained refinement are easily understood by considering a *one-dimensional protein consisting of two like atoms* in the asymmetric unit, with coordinates x_1 and x_2 relative to a fixed origin and bond length $l = x_2 - x_1$. In the refinement, the normal equations are of the type $\mathbf{N}\Delta\mathbf{x} = \mathbf{e}$. For two non-overlapping like atoms, the *diffraction data* will yield a normal matrix

$$\mathbf{N} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \qquad (18.5.3.3)$$

with inverse

$$\begin{pmatrix} 1/a & 0 \\ 0 & 1/a \end{pmatrix}, \qquad (18.5.3.4)$$

where

$$a = \sum w_h(\partial|F_n|/\partial x_i)^2. \qquad (18.5.3.5)$$

A *geometric restraint* on the length will yield a normal matrix

$$\begin{pmatrix} b & -b \\ -b & b \end{pmatrix} \qquad (18.5.3.6)$$

with no inverse, since its determinant is zero, where

$$b = w_{\rm geom}(\partial l/\partial x_i)^2. \qquad (18.5.3.7)$$

Note $\partial l/\partial x_2 = -\partial l/\partial x_i = 1$, so that

$$b = w_{\rm geom} = 1/\sigma_{\rm geom}^2(l), \qquad (18.5.3.8)$$

where $\sigma^2_{\rm geom}(l)$ is the variance assigned to the length in the stereochemical dictionary.

Combining the diffraction data and the restraint, the normal matrix becomes

$$\begin{pmatrix} a+b & -b \\ -b & a+b \end{pmatrix}, \tag{18.5.3.9}$$

with inverse

$$\{1/[a(a+2b)]\}\begin{pmatrix} a+b & b \\ b & a+b \end{pmatrix}. \tag{18.5.3.10}$$

For the diffraction data alone, the variance of x_i is

$$\sigma^2_{\rm diff}(x_i) = 1/a. \tag{18.5.3.11}$$

For the diffraction data plus restraint, the variance of x_i is

$$\begin{aligned}\sigma^2_{\rm res}(x_i) &= (a+b)/[a(a+2b)]\\ &< \sigma^2_{\rm diff}(x_i).\end{aligned} \tag{18.5.3.12}$$

Note that though the restraint says nothing about the position of x_i, the variance of x_i has been reduced because of the coupling to the position of the other atom. In the limit when $a \ll b$, $\sigma^2_{\rm res}(x_i)$ is only half $\sigma^2_{\rm diff}(x_i)$.

The general formula for the variance of the length $l = x_2 - x_1$ is

$$\sigma^2(l) = \sigma^2(x_2) - 2{\rm cov}(x_2, x_1) + \sigma^2(x_1). \tag{18.5.3.13}$$

For the diffraction data alone, this gives

$$\sigma^2_{\rm diff}(l) = 1/a + 0 + 1/a = 2/a = 2\sigma^2_{\rm diff}(x_i), \tag{18.5.3.14}$$

as expected. For the diffraction data plus restraint,

$$\begin{aligned}\sigma^2_{\rm res}(l) &= [1/a(a+2b)][(a+b) - 2b + (a+b)]\\ &= 1/(a/2+b)\\ &< \sigma^2_{\rm diff}(l).\end{aligned} \tag{18.5.3.15}$$

For small a, $\sigma^2_{\rm res}(l) \to 1/b = \sigma^2_{\rm geom}(l)$, as expected. The variance of the restrained length, (18.5.3.15), can be re-expressed as

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l). \tag{18.5.3.16}$$

For the two-atom protein, it can be proved directly, as one would expect from (18.5.3.16), that *restrained refinement determines a length which is the weighted mean of the diffraction-only length and the geometric dictionary length.*

The centroid has coordinate $c = (x_1 + x_2)/2$. It is easily found that $\sigma^2_{\rm res}(c) = \sigma^2_{\rm diff}(c) = 1/2a$. Thus, as expected, the restraint says nothing about the position of the molecule in the cell.

For numerical illustrations of the s.u.'s in restrained refinement, suppose the stereochemical length restraint has $\sigma_{\rm geom}(l) = 0.02$ Å. Equation (18.5.3.16) gives the length s.u. $\sigma_{\rm res}(l)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(l)$ is 0.012 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(l)$ is 0.019 Å. However large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(l)$ never exceeds 0.02 Å.

Equation (18.5.3.12) gives the position s.u. $\sigma_{\rm res}(x_i)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(x_i)$ is 0.009 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(x_i) = 0.037$ Å. For large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(x_i)$ tends to $\sigma_{\rm diff}(x_i)/(2)^{1/2}$ as the strong restraint couples the two atoms together. For very small $\sigma_{\rm diff}(x_i)$, the relatively weak restraint has no effect.

18.5.3.3. *Relative weighting of diffraction and restraint terms*

When only relative diffraction weights are known, as in equation (18.5.2.13), it has been common (Rollett, 1970) to scale the geometric restraint terms against the diffraction terms by replacing the restraint weights $w_{\rm geom} = 1/\sigma^2_{\rm geom}$ by $w_{\rm geom} = S^2/\sigma^2_{\rm geom}$, where $S^2 = (\sum w_h \Delta_h^2)/(n_{\rm obs} - n_{\rm params})$. However, this scheme cannot be used for low-resolution structures if $n_{\rm obs} < n_{\rm params}$.

The treatment by Tickle *et al.* (1998*a*) shows that the reduction $n_{\rm params}$ in the number of degrees of freedom has to be distributed among all the data, both diffraction observations and restraints. Since the geometric restraint weights are on an absolute scale (Å^{-2}), they propose that the (absolute) scale of the diffraction weights should be determined by adjustment until the restrained residual R' (18.5.3.1) is equal to its expected value ($n_{\rm obs} + n_{\rm restraints} - n_{\rm params}$).

For a method of determining the scale of the diffraction weights based on $R_{\rm free}$, see Brünger (1993).

The geometric restraint weights were classified by the IUCr Subcommittee (Schwarzenbach *et al.*, 1995) as derived from observations supplementary to the diffraction data, with uncertainties of type B (Section 18.5.2.3).

18.5.4. Two examples of full-matrix inversion

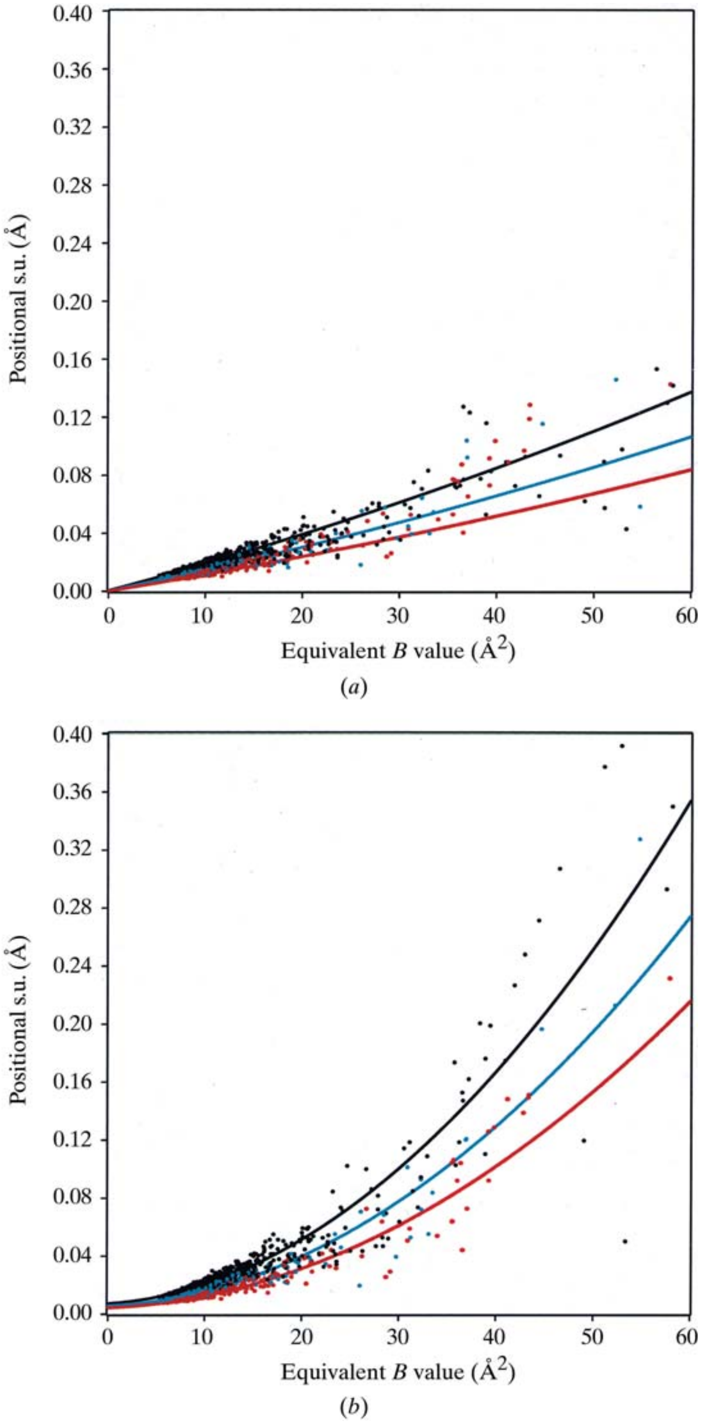
18.5.4.1. *Unrestrained and restrained inversions for concanavalin A*

G. M. Sheldrick extended his *SHELXL*96 program (Sheldrick & Schneider, 1997) to provide extra information about protein precision through the inversion of least-squares full matrices. His programs have been used by Deacon *et al.* (1997) for the high-resolution refinement of native concanavalin A with 237 residues, using data at 110 K to 0.94 Å refined anisotropically. After the convergence and completion of full-matrix restrained refinement for the structure, the unrestrained full matrix (coordinates only) was computed and then inverted in a massive calculation. This led to s.u.'s $\sigma(x)$, $\sigma(y)$, $\sigma(z)$ and $\sigma(r)$ for all atoms, and to $\sigma(l)$ and $\sigma(\theta)$ for all bond lengths and angles. $\sigma(r)$ is defined as $[\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}$. For concanavalin A the restrained full matrix was also inverted, thus allowing the comparison of restrained and unrestrained s.u.'s.

The results for concanavalin A from the inversion of the coordinate matrices of order 6402 (= 2134 × 3) are plotted in Figs. 18.5.4.1 and 18.5.4.2. Fig. 18.5.4.1 shows $\sigma(r)$ *versus* $B_{\rm eq}$ for the fully occupied atoms of the protein (a few atoms with $B >$ 60 Å^2 are off-scale). The points are colour-coded black for carbon, blue for nitrogen and red for oxygen. Fig. 18.5.4.1(*a*) shows the restrained results, and Fig. 18.5.4.1(*b*) shows the unrestrained diffraction-data-only results. Superposed on both sets of data points are least-squares quadratic fits determined with weights $1/B^2$. At high B, the unrestrained $\sigma_{\rm diff}(r)$ can be at least double the restrained $\sigma_{\rm res}(r)$, *e.g.*, for carbon at B = 50 Å^2, the unrestrained $\sigma_{\rm diff}(r)$ is about 0.25 Å, whereas the restrained $\sigma_{\rm res}(r)$ is about 0.11 Å. For B < 10 Å^2, both $\sigma(r)$'s fall below 0.02 Å and are around 0.01 Å at B = 6 Å^2.

For B < 10 Å^2, the better precision of oxygen as compared with nitrogen, and of nitrogen as compared with carbon, can be clearly seen. At the lowest B, the unrestrained $\sigma_{\rm diff}(r)$ in Fig. 18.5.4.1(*b*) are almost as small as the restrained $\sigma_{\rm res}(r)$ in Fig. 18.5.4.1(*a*). [The quadratic fits of the restrained results in Fig. 18.5.4.1(*a*) are evidently slightly imperfect in making $\sigma_{\rm res}(r)$ tend almost to 0 as B tends to 0.]

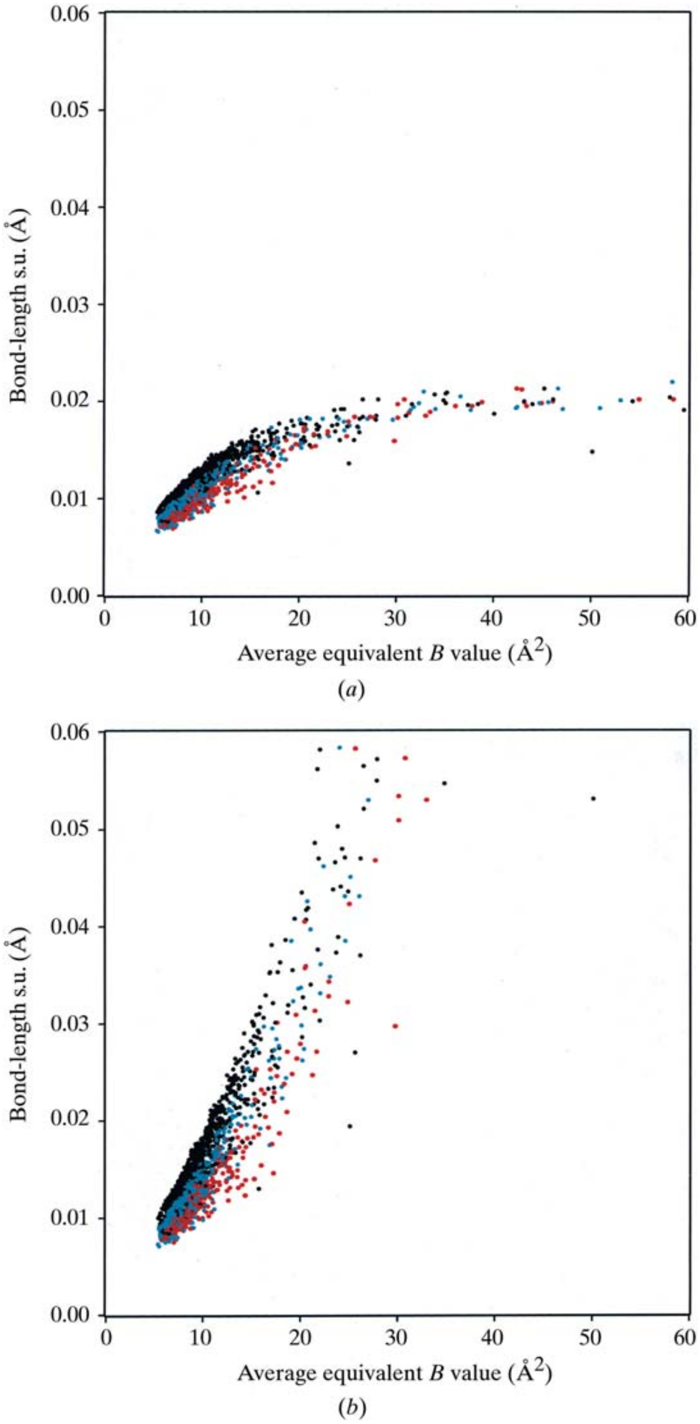
Fig. 18.5.4.2 shows $\sigma(l)$ *versus* $B_{\rm eq}$ for the bond lengths in the protein. The points are colour-coded black for C—C, blue for C—N and red for C—O. The restrained and unrestrained distributions are very different for high B. The restrained distribution in



(*a*)

(*b*)

Figure 18.5.4.1
Plots of $\sigma(r)$ *versus* $B_{\rm eq}$ for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{\rm res}(r)$, (*b*) unrestrained full-matrix $\sigma_{\rm diff}(r)$. Carbon black, nitrogen blue, oxygen red.



(*a*)

(*b*)

Figure 18.5.4.2
Plots of $\sigma(l)$ *versus* average $B_{\rm eq}$ for concanavalin A with 0.94 Å data, (*a*) restrained full-matrix $\sigma_{\rm res}(l)$, (*b*) unrestrained full-matrix $\sigma_{\rm diff}(l)$. C—C black, C—N blue, C—O red.

Fig. 18.5.4.2(*a*) tends to about 0.02 Å, which is the standard uncertainty of the applied restraint for 1–2 bond lengths, whereas the unrestrained distribution in Fig. 18.5.4.2(*b*) goes off the scale of the diagram. But for $B < 10$ Å^2, both distributions fall to around 0.01 Å.

The differences between the restrained and unrestrained $\sigma(r)$ and $\sigma(l)$ can be understood through the two-atom model for restrained refinement described in Section 18.5.3. For that model, the equation

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l) \qquad (18.5.3.16)$$

relates the bond-length s.u. in the restrained refinement, $\sigma_{\rm res}(l)$, to the $\sigma_{\rm diff}(l)$ of the unrestrained refinement and the s.u. $\sigma_{\rm geom}(l)$ assigned to the length in the stereochemical dictionary. In the refinements, $\sigma_{\rm geom}(l)$ was 0.02 Å for all bond lengths. When this is combined in (18.5.3.16) with the unrestrained $\sigma_{\rm diff}(l)$ of any bond, the predicted restrained $\sigma_{\rm res}(l)$ is close to that found in the restrained full matrix.

It can be seen from Fig. 18.5.4.2(*b*) that many bond lengths with average $B < 10$ Å^2 have $\sigma_{\rm diff}(l) < 0.014$ Å. For these bonds the diffraction data have greater weight than the stereochemical dictionary. Some bonds have $\sigma_{\rm diff}(l)$ as low as 0.0080 Å, with $\sigma_{\rm res}(l)$ around 0.0074 Å. This situation is one consequence of the availability of diffraction data to the high resolution of 0.94 Å. For large $\sigma_{\rm diff}(l)$ (*i.e.*, high B), equation (18.5.3.16)

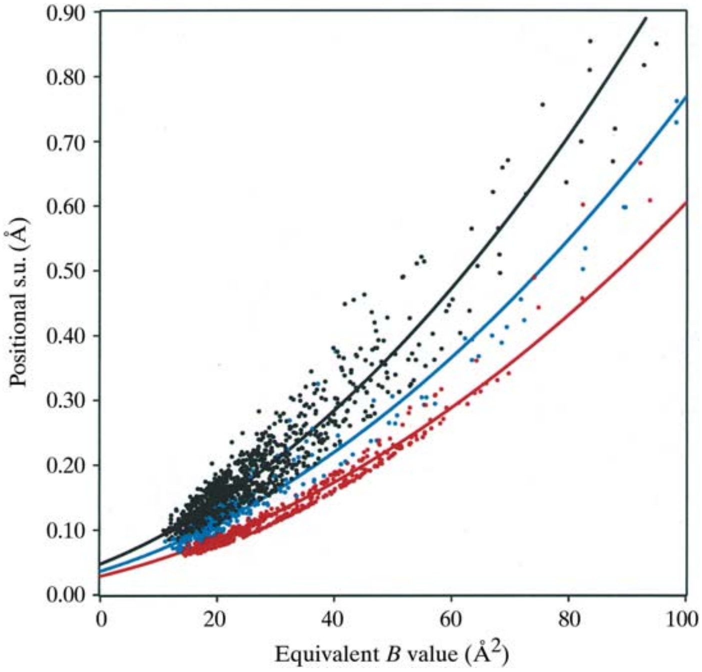


Figure 18.5.4.3
Plot of $\sigma_{\rm diff}(r)$ *versus* $B_{\rm eq}$ from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. Carbon black, nitrogen blue, oxygen red.

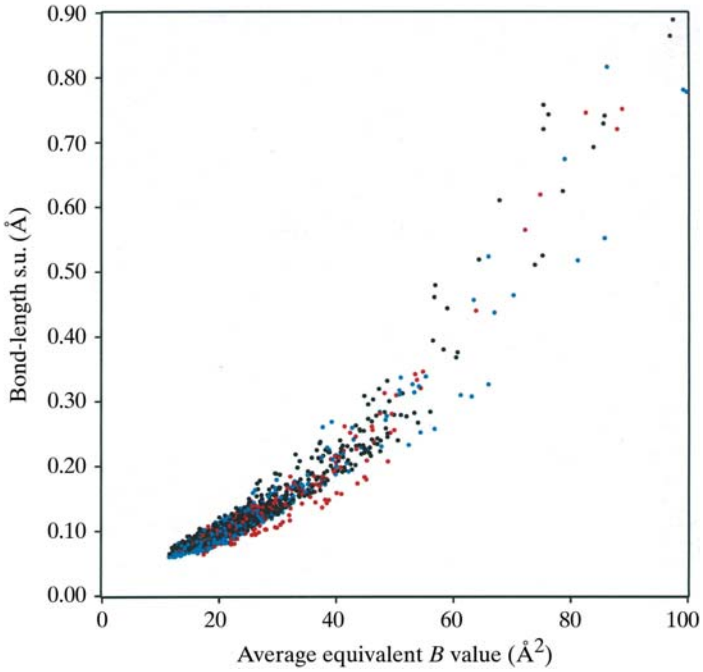


Figure 18.5.4.4
Plot of $\sigma_{\rm diff}(l)$ *versus* average $B_{\rm eq}$ from an unrestrained full matrix for immunoglobulin mutant (T39K) with 1.70 Å data. C—O black, C—N blue, C—O red.

predicts that $\sigma_{\rm res}(l) = \sigma_{\rm geom}(l) = 0.02$ Å, as is found in Fig. 18.5.4.2(*a*).

In an isotropic approximation, $\sigma(r) = 3^{1/2}\sigma(x)$. Equation (18.5.3.12) of the two-atom model can be recast to give

$$\sigma^2_{\rm res}(r) = \sigma^2_{\rm diff}(r)\left\{\left[\sigma^2_{\rm diff}(r) + 3(0.02)^2\right]\Big/\left[2\sigma^2_{\rm diff}(r) + 3(0.02)^2\right]\right\}. \qquad (18.5.4.1)$$

For low B, say $B \leq 15$ Å^2 in concanavalin, (18.5.4.1) gives quite good predictions of $\sigma_{\rm res}(r)$ from $\sigma_{\rm diff}(r)$. For instance, for a carbon atom with B = 15 Å^2, the quadratic curve for carbon in Fig. 18.5.4.1(*b*) shows $\sigma_{\rm diff}(r) = 0.034$ Å, and Fig. 18.5.4.1(*a*) shows $\sigma_{\rm res}(r) = 0.029$ Å. While if $\sigma_{\rm diff}(r) = 0.034$ Å is used with (18.5.4.1), the resulting prediction for $\sigma_{\rm res}(r)$ is 0.028 Å.

However, for high B, say B = 50 Å^2, the quadratic curve for carbon in Fig. 18.5.4.1(*b*) shows $\sigma_{\rm diff}(r) = 0.25$ Å, and Fig. 18.5.4.1(*a*) shows $\sigma_{\rm res}(r) = 0.11$ Å, whereas (18.5.4.1) leads to the poor estimate $\sigma_{\rm res}(r) = 0.18$ Å.

Thus at high B, equation (18.5.4.1) from the two-atom model does not give a good description of the relationship between the restrained and unrestrained $\sigma(r)$. The reason is obvious. Most atoms are linked by 1–2 bond restraints to two or three other atoms. Even a carbonyl oxygen atom linked to its carbon atom by a 0.02 Å restraint is also subject to 0.04 Å 1–3 restraints to chain C_α and N atoms. Consequently, for a high-B atom, when the restraints are applied it is coupled to several other atoms in a group, and its $\sigma_{\rm res}(r)$ is lower, compared with the diffraction-data-only $\sigma_{\rm diff}(r)$, by a greater amount than would be expected from the two-atom model.

18.5.4.2. *Unrestrained inversion for an immunoglobulin*

Sheldrick has provided the results of the unrestrained lower-resolution refinement of a single-chain immunoglobulin mutant (T39K) with 218 amino-acid residues, with data to 1.70 Å refined isotropically (Usón *et al.*, 1999). Fig. 18.5.4.3 shows $\sigma_{\rm diff}(r)$ *versus* $B_{\rm eq}$ for the fully occupied protein atoms. Superposed on the data points are least-squares quadratic fits. In a first very rough approximation for $\sigma_{\rm diff}(x_i)$ suggested later by equation (18.5.6.3), the dependence on atom type is controlled by $1/Z_i$, the reciprocal of the atomic number. Sheldrick found that a $1/Z_i$ dependence produced too little difference between C, N and O. The proportionalities between the quadratics for $\sigma(r)$ in Figs. 18.5.4.1 and 18.5.4.3 are based on the reciprocals of the scattering factors at $\sin\theta/\lambda = 0.3$ Å^{-1}, symbolized by $Z^\#_i$. For C, N and O, these are 2.494, 3.219 and 4.089, respectively. For potential use in later work, the least-squares fits to the $\sigma(r_i)Z^\#_i$ in Å are recorded here as

$$0.11892 + 0.00891B + 0.0001462B^2, \qquad (18.5.4.2a)$$

$$0.01826 + 0.001043B + 0.0002230B^2 \text{ and} \qquad (18.5.4.2b)$$

$$0.00115 + 0.004414B + 0.0000214B^2 \qquad (18.5.4.2c)$$

for the immunoglobulin (unrestrained), concanavalin A (unrestrained) and concanavalin A (restrained), respectively.

As might be expected from the lower resolution, the lowest $\sigma_{\rm diff}(r)$'s in the immunoglobulin are about six times the lowest $\sigma_{\rm diff}(r)$'s in concanavalin. But at B = 50 Å^2, the immunoglobulin curve for carbon gives $\sigma_{\rm diff}(r) = 0.37$ Å, which is only 50% larger than the concanavalin value of 0.25 Å.

Fig. 18.5.4.4 shows $\sigma_{\rm diff}(l)$ *versus* $B_{\rm eq}$ for the immunoglobulin. Note that the lowest immunoglobulin unrestrained $\sigma_{\rm diff}(l)$ is about 0.06 Å, which is three times the 0.02 Å $\sigma_{\rm geom}(l)$ bond restraint.

18.5.4.3. *Comments on restrained refinement*

Geometric restraint dictionaries typically use bond-length weights based on $\sigma_{\rm geom}(l)$ of around 0.02 or 0.03 Å. Tables 18.5.7.1–18.5.7.3 show that even 1.5 Å studies have diffraction-only errors $\sigma_{\rm diff}(x, B_{\rm avg})$ of 0.08 Å and upwards. Only for resolutions of 1.0 Å or so are the diffraction-only errors comparable with the dictionary weights. Of course, the dictionary offers no values for many of the configurational parameters of the protein structure, including the centroid and molecular orientation.

18.5.4.4. Full-matrix estimates of precision

The opening contention of this chapter in Section 18.5.1.1 is that the variances and covariances of the structural parameters of proteins can be found from the inverse of the least-squares normal matrix. But there is a caveat, chiefly that explicit account would not be taken of disorder of the solvent or of parts of the protein. Corrections by Babinet's principle of complementarity or by mask bulk solvent models are only first-order approximations. The consequences of such disorder problems, which make the variation of calculated structure factors nonlinear over the range of interest, may in future be better handled by maximum-likelihood methods (*e.g.* Read, 1990; Bricogne, 1993; Bricogne & Irwin, 1996; Murshudov *et al.*, 1997). Pannu & Read (1996) have shown how the maximum-likelihood method can be cast computationally into a form akin to least-squares calculations. Full-matrix precision estimates along the lines of the present chapter are probably somewhat low.

It should also be noted that full-matrix estimates of coordinate precision are most reliably derived from matrices involving both coordinates and atomic displacement parameters. This is particularly important for lower-resolution analyses, in which atomic images overlap. The work on the high-resolution analysis of concanavalin A described in Section 18.5.4.1 was based on the very large coordinate matrix, of order 6402. The omission, because of computer limitations, of the anisotropic displacement parameters from the full matrix will have caused the coordinate s.u.'s of atoms with high $B_{\rm eq}$ to be underestimated.

Much information about the quality of a molecular model can be obtained from the eigenvalues and eigenvectors of the normal matrix (Cowtan & Ten Eyck, 2000).

18.5.5. Approximate methods

18.5.5.1. Block calculations

The full-matrix inversions described in the previous section require massive calculations. The length of the calculations is more a matter of the order of the matrix, *i.e.*, the number of parameters, than of the number of observations. When restraints are applied, it is the diffraction-cum-restraints full matrix which should be inverted.

With the increasing power of computers and more efficient algorithms (*e.g.* Tronrud, 1999; Murshudov *et al.*, 1999), a final full matrix should be computed and inverted much more regularly – and not just for high-resolution analyses. Low-resolution analyses have a need, beyond the indications given by B values, to identify through $\sigma(x)$ estimates their regions of tolerable and less tolerable precision.

If full-matrix calculations are impractical, partial schemes can be suggested. As far back as 1973, Watenpaugh *et al.* (1973), in a study of rubredoxin at 1.5 Å resolution, effectively inverted the diffraction full matrix in 200 parameter blocks to obtain individual s.u.'s. A similar scheme for restrained refinements could also use overlapping large blocks. A minimal block scheme in refinements of any resolution is to calculate blocks for each residue and for the block interactions between successive residues. The inversion process could then use the matrices in running groups of three successive residues, taking only the inverted elements for the central residue as the estimates of its variances and covariances.

For low-resolution analyses with very large numbers of atoms, it might be sufficient to gain a general idea of the behaviour of $\sigma(x)$ as a function of B by computing a limited number of blocks for representative or critical groups of residues. The parameters used in the blocks should include the B's, since atomic images overlap at low resolution, thus correlating the position of one atom with the displacement parameters of its neighbours.

18.5.5.2. The modified Fourier method

In the simplest form of the Fourier-map approach to centrosymmetric high-resolution structures, atomic positions are given by the maxima of the observed electron density. The uncertainty of such a position may be estimated as the uncertainty in the slope function (first derivative) divided by the curvature (second derivative) at the peak (Cruickshank, 1949*a*), *i.e.*,

$$\sigma(x) = \sigma(\text{slope})/(\text{atomic peak `curvature'}). \qquad (18.5.5.1)$$

However, atomic positions are affected by finite-series and peak-overlapping effects.

Hence, more generally, atomic positions may be determined by the requirement that the slope of the difference map at the position of atom r should be zero, or equivalently that the slopes at atom r of the observed and calculated electron densities should be equal. As a criterion this becomes the basis of the modified Fourier method (Cruickshank, 1952, 1959, 1999; Bricogne, 2008, Section 1.3.4.4.7.5), which, like the least-squares method, is applicable whether or not the atomic peaks are resolved and is applicable to noncentrosymmetric structures. For refinement, a set of n simultaneous linear equations are involved, analogous to the normal equations of least squares. Their right-hand sides are the slopes of the difference map at the trial atomic positions.

The diagonal elements of the matrix, for coordinate x_r of an atom with Debye B value B_r, are approximately equal to

$$\text{`curvature'} = (4\pi^2/a^2V)\left[\sum_{hkl}(m/2)h^2 f_r \exp(-B_r \sin^2\theta/\lambda^2)\right], \qquad (18.5.5.2)$$

where $m = 1$ or 2 for acentric or centric reflections. The summation is over all independent planes and their symmetry equivalents. Strictly speaking, (18.5.5.2) is a curvature only for centrosymmetric structures.

In the modified Fourier method,

$$\sigma(\text{slope}) = (2\pi/aV)\left[\sum_{hkl} h^2(\Delta|F|^2)\right]^{1/2}. \qquad (18.5.5.3)$$

This is simply an estimate of the r.m.s. uncertainty at a general position (Cruickshank & Rollett, 1953) in the slope of the difference map, *i.e.*, the r.m.s. uncertainty on the right-hand side of the modified Fourier method.

$\sigma(x)$ is then given by (18.5.5.1), using (18.5.5.3) and (18.5.5.2).

18.5.5.3. Application of the modified Fourier method

An extreme example of an apparently successful gross approximation to protein precision is represented by Daopin *et al.*'s (1994) treatment of two independent determinations (at 1.8 and 1.95 Å) of the structure of TGF-β2. They reported that the modified Fourier-map formulae given in Section 18.5.5.2 yielded a quite good description of the B dependence of the positional differences between the two independent determinations. However, there is a formal difficulty about this application. Equation (18.5.5.1) derives from a diffraction-data-only approach, whereas the two structures were determined from restrained refinements. Even though the *TNT* restraint parameters and weights may have been the same in both refinements, it is slightly surprising that (18.5.5.1) should have worked well.

Equation (18.5.2.1) requires the summation of various series over all (*hkl*) observations; such calculations are not customarily provided in protein programs. However, due to the fundamental similarities between Fourier and least-squares methods demonstrated by Cochran (1948), Cruickshank (1949*b*, 1952, 1959), and Cruickshank & Robertson (1953), closely similar estimates of the precision of individual atoms can be obtained from the reciprocal of the diagonal elements of the diffraction-data-only least-squares matrix. These elements will often have been calculated already within the protein refinement programs, but possibly never output. Such estimates could be routinely available.

Between approximations using largish blocks and those using only the reciprocals of diagonal terms, a whole variety of intermediate approximations involving some off-diagonal terms could be envisaged.

Whatever method is used to estimate uncertainties, it is essential to distinguish between *coordinate* uncertainty, *e.g.*, $\sigma(x)$, and *position* uncertainty $\sigma(r) = [\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}$.

The remainder of this chapter discusses two rough-and-ready indicators of structure precision: the diffraction-component precision index (DPI) and Luzzati plots.

18.5.6. The diffraction-component precision index

18.5.6.1. Statistical expectation of error dependence

From general statistical theory, one would expect the s.u. of an atomic coordinate determined from the diffraction data alone to show dependence on four factors:

$$\sigma(x) \propto (\mathcal{R})\big[(n_{\rm atoms})/(n_{\rm obs} - n_{\rm params})\big]^{1/2}(1/s_{\rm rms}). \qquad (18.5.6.1)$$

Here, $\mathcal{R}$ is some measure of the precision of the data; $n_{\rm atoms}$ is the recognition that the information content of the data has to be shared out; $n_{\rm obs}$ is the number of independent data, but to achieve the correct number of degrees of freedom this must be reduced by $n_{\rm params}$, the number of parameters determined; and $1/s_{\rm rms}$ is a more specialized factor arising from the sensitivity $\partial|F|/\partial x$ of the data to the parameter x. Here $s_{\rm rms}$ is the r.m.s. reciprocal radius of the data. Any statistical error estimate must show some correspondence to these four factors.

18.5.6.2. A simple error formula

Cruickshank (1960) offered a simple order-of-magnitude formula for $\sigma(x)$ in small molecules. It was intended for use in experimental design: how many data of what precision are needed to achieve a given precision in the results? The formula, derived from a very rough estimate of a least-squares diagonal element in non-centrosymmetric space groups, was

$$\sigma(x_i) = (1/2)(N_i/p)^{1/2}[R/s_{\rm rms}]. \qquad (18.5.6.2)$$

Here $p = n_{\rm obs} - n_{\rm params}$, R is the usual residual $\sum|\Delta F|/\sum|F|$ and N_i is the number of atoms of type i needed to give scattering power at $s_{\rm rms}$ equal to that of the asymmetric unit of the structure, *i.e.*, $\sum_j f_j^2 \equiv N_i f_i^2$. [The formula has also proved very useful in a systematic study of coordinate precision in the many thousands of small-molecule structure analyses recorded in the Cambridge Structural Database (Allen *et al.*, 1995*a*,*b*).]

For small molecules, the above definition of N_i allowed the treatment of different types of atom with not-too-different B's. However, it is not suitable for individual atoms in proteins where there is a very large range of B values and some atoms have B's so large as to possess negligible scattering power at $s_{\rm rms}$.

Often, as in isotropic refinement, $n_{\rm params} \simeq 4n_{\rm atoms}$, where $n_{\rm atoms}$ is the total number of atoms in the asymmetric unit. For fully anisotropic refinement, $n_{\rm params} \simeq 9n_{\rm atoms}$.

A first very rough extension of (18.5.6.2) for application in proteins to an atom with $B = B_i$ is

$$\sigma(x_i) = k(N_i/p)^{1/2}\big[g(B_i)/g(B_{\rm avg})\big]C^{-1/3}Rd_{\rm min}, \qquad (18.5.6.3)$$

where k is about 1.0, $N_i = \sum Z_j^2/Z_i^2$, $B_{\rm avg}$ is the average B for fully occupied sites and C is the fractional completeness of the data to $d_{\rm min}$. In deriving (18.5.6.3) from (18.5.6.2), $1/s_{\rm rms}$ has been replaced by $1.3d_{\rm min}$, and the factor $(1/2)(1.3) = 0.65$ has been increased to 1.0 as a measure of caution in the replacement of a full matrix by a diagonal approximation. $g(B) = 1 + a_1B + a_2B^2$ is an empirical function to allow for the dependence of $\sigma(x)$ on B. However, the results in Section 18.5.4.2 showed that the parameters a_1 and a_2 depend on the structure.

As also mentioned in Section 18.5.4.2, Sheldrick has found that the Z_i in N_i is better replaced by $Z_i^{\#}$, the scattering factor at $\sin\theta/\lambda = 0.3$ Å^{-1}. Hence, N_i may be taken as

$$N_i = (\textstyle\sum Z_j^{\#2}/Z_i^{\#2}). \qquad (18.5.6.4)$$

A useful comparison of the relative precision of different structures may be obtained by comparing atoms with the respective $B = B_{\rm avg}$ in the different structures. (18.5.6.3) then reduces to

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/p)^{1/2}C^{-1/3}Rd_{\rm min}. \qquad (18.5.6.5)$$

The smaller the $d_{\rm min}$ and the R, the better the precision of the structure. If the difference between oxygen, nitrogen and carbon atoms is ignored, N_i may be taken simply as the number of fully occupied sites. For heavy atoms, (18.5.6.4) must be used for N_i.

Equation (18.5.6.5) is not to be regarded as having absolute validity. It is a quick and rough guide for the diffraction-data-only error component for an atom with Debye B equal to the $B_{\rm avg}$ for the structure. It is named the *diffraction-component precision index*, or DPI. It contains none of the restraint data.

18.5.6.3. Extension for low-resolution structures and use of R_{free}

For low-resolution structures, the number of parameters may exceed the number of diffraction data. In (18.5.6.3) and (18.5.6.5), $p = n_{\rm obs} - n_{\rm params}$ is then negative, so that $\sigma(x)$ is imaginary. This difficulty can be circumvented *empirically* by replacing p with $n_{\rm obs}$ and R with $R_{\rm free}$ (Brünger, 1992). The counterpart of the DPI (18.5.6.5) is then

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min}. \qquad (18.5.6.6)$$

Here $n_{\rm obs}$ is the number of reflections included in the refinement, not the number in the $R_{\rm free}$ set.

It may be asked: how can there be any estimate for the precision of a coordinate from the diffraction data only when there are insufficient diffraction data to determine the structure? By following the line of argument of Cruickshank's (1960) analysis, (18.5.6.6) is a rough estimate of the square root of the reciprocal of one diagonal element of the diffraction-only least-squares matrix. All the other parameters can be regarded as having been determined from a diffraction-plus-restraints matrix.

Clearly, (18.5.6.6) can also be used as a general alternative to (18.5.6.5) as a DPI, irrespective of whether the number of degrees of freedom $p = n_{\rm obs} - n_{\rm params}$ is positive or negative.

Table 18.5.7.1
Comparison of full-matrix $\sigma(r, B_{\rm avg})$ with the diffraction-component precision index (DPI)

Protein	$(N_i/p)^{1/2}$	R	$d_{\min}$ (Å)	DPI $\sigma(r, B_{\rm avg})$ (Å)	Full-matrix $\sigma_{\rm diff}(r, B_{\rm avg})$ (Å)	Reference
Concanavalin A	0.148	0.128	0.94	0.034	0.033	(*a*)
Immunoglobulin	0.476	0.156	1.70	0.221	0.186	(*b*)

References: (*a*) Deacon *et al.* (1997); (*b*) Usón *et al.* (1999).

Comment. When p is positive, (18.5.6.6) would be exactly equivalent to (18.5.6.5) only if $R_{\rm free} = R[n_{\rm obs}/(n_{\rm obs} - n_{\rm params})]^{1/2}$. Tickle *et al.* (1998*b*) have shown that the expected relationship in a restrained refinement is actually

$$R_{\rm free} = R\{[n_{\rm obs} + (n_{\rm params} - h)]/[n_{\rm obs} - (n_{\rm params} - h)]\}^{1/2}, \qquad (18.5.6.7)$$

where $h = n_{\rm restraints} - \sum w_{\rm geom}(\Delta Q)^2$, the latter term, as in (18.5.3.1), being the weighted sum of the squares of the restraint residuals.

18.5.6.4. Position error

Often an estimate of a position error $|\Delta \mathbf{r}|$, rather than a coordinate error $|\Delta x|$, is required. In the isotropic approximation,

$$\sigma(r, B_{\rm avg}) = 3^{1/2}\sigma(x, B_{\rm avg}). \qquad (18.5.6.8)$$

Consequently, the DPI formulae for the position errors are

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/p)^{1/2}C^{-1/3}Rd_{\min} \qquad (18.5.6.9)$$

with R and

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\min} \qquad (18.5.6.10)$$

with $R_{\rm free}$.

18.5.7. Examples of the diffraction-component precision index

18.5.7.1. Full-matrix comparison with the diffraction-component precision index

The DPI (18.5.6.9) with R was offered as a quick and rough guide for the diffraction-data-only error for an atom with $B = B_{\rm avg}$. The necessary data for the comparison with the two unrestrained full-matrix inversions of Section 18.5.5 are given in Table 18.5.7.1. For concanavalin A with $B_{\rm avg} = 14.8$ Å^2, the full-matrix quadratic (18.5.4.2*b*) gives 0.033 Å for a carbon atom and the DPI gives 0.034 Å for an unspecified atom. For the immunoglobulin with $B_{\rm avg} = 26.8$ Å^2, the full-matrix quadratic (18.5.4.2*a*) gives $\sigma_{\rm diff}(r) = 0.19$ Å for a carbon atom, while the DPI gives 0.22 Å.

For these two structures, the simple DPI formula compares surprisingly well with the unrestrained full-matrix calculations at $B_{\rm avg}$.

For the restrained full-matrix calculations on concanavalin A, the quadratic (18.5.4.2*c*) with $B = B_{\rm avg}$ gives $\sigma_{\rm res}(r) = 0.028$ Å for a carbon atom, which is only 15% smaller than the unrestrained 0.033 Å. This small decrease matches the discussion of $\sigma_{\rm res}(r)$ and $\sigma_{\rm diff}(r)$ in Section 18.5.4.1 following equation (18.5.4.1). But that discussion also indicates that for the immunoglobulin, the restrained $\sigma_{\rm res}(r, B_{\rm avg})$, which was not computed, will be proportionaly much lower than the unrestrained value of $\sigma_{\rm diff}(r, B_{\rm avg}) = 0.19$ Å, since the restraints are relatively more important in the immunoglobulin.

18.5.7.2. Further examples of the DPI using R

Table 18.5.7.2 shows a range of examples of the application of the DPI (18.5.6.9) using R to proteins of differing precision, starting with the smallest $d_{\min}$. In all the examples, N_i has been set equal to $n_{\rm atoms}$, the total number of atoms. The ninth and tenth columns show $\langle \Delta r\rangle$ values derived from Luzzati (1952) and Read (1986) plots described later in Section 18.5.8.

The first entry is for crambin at 0.83 Å resolution and 130 K (Stec *et al.*, 1995). Their results were obtained from an unrestrained full-matrix anisotropic refinement. Inversion of the full matrix gave s.u.'s $\sigma_{\rm diff}(x) = 0.0096$ Å for backbone atoms, 0.0168 Å for side-chain atoms and 0.0409 Å for solvent atoms, with an average for all atoms of 0.022 Å. The DPI $\sigma(r, B_{\rm avg}) = 0.021$ Å corresponds to $\sigma(x) = 0.012$ Å, which is satisfactorily intermediate between the full-matrix values for the backbone and side-chain atoms.

Sevcik *et al.* (1996) carried out restrained anisotropic full-matrix refinements on data from two slightly different crystals of ribonuclease Sa, with $d_{\min}$ of 1.15 and 1.20 Å. They inverted full-matrix blocks containing parameters of 20 residues to estimate coordinate errors. The overall r.m.s. coordinate error for protein atoms is given as 0.03 Å, and for all atoms (including waters and ligands) as 0.07 Å for MGMP and 0.05 Å for MSA. The DPI gives $\sigma(r, B_{\rm avg}) = 0.05$ Å for both structures.

The next entries concern the two lower-resolution (1.8 and 1.95 Å) studies of TGF-β2 (Daopin *et al.*, 1994). The DPI gives $\sigma(r) = 0.16$ Å for 1TGI and 0.24 Å for 1TGF. This indicates an

Table 18.5.7.2
Examples of diffraction-component precision indices (DPIs)

Protein	N_i	$n_{\rm obs}$	$(N_i/p)^{1/2}$	$C^{-1/3}$	R	$d_{\min}$ (Å)	DPI $\sigma(r, B_{\rm avg})$ (Å)	Luzzati $\langle \Delta r\rangle$ (Å)	Read $\langle \Delta r\rangle$ (Å)	Reference
Crambin	447	23759	0.150	1.074	0.090	0.83	0.021	0.055		(*a*)
Ribonuclease MGMP	1958	62845	0.208	1.046	0.109	1.15	0.047		0.08	(*b*)
Ribonuclease MSA	1832	60670	0.204	1.016	0.106	1.20	0.045		0.05	(*b*)
TGF-β2 1TGI	948	~14000	0.305	~1.0	0.173	1.80	0.16	0.21	0.18	(*c*)
TGF-β2 1TFG	974	~11000	0.370	~1.0	0.188	1.95	0.24	0.23		(*c*)
Lactoferrin	5907	39113	0.618	1.036	0.179	2.20	0.43	0.25–0.30	0.35	(*d*)
Thaumatin C2	1552	4622	†	1.10	0.184	2.60	—	0.25		(*e*)

References: (*a*) Stec *et al.* (1995); (*b*) Sevcik *et al.* (1996); (*c*) Daopin *et al.* (1994); (*d*) Haridas *et al.* (1995); (*e*) Ko *et al.* (1994). † (N_i/p) negative.

Table 18.5.7.3
Comparison of DPIs using R and $R_{\rm free}$

The second row for each protein contains values appropriate to the DPI equation (18.5.6.10) using $R_{\rm free}$.

Protein	N_i	$n_{\rm obs}$	$(N_i/p)^{1/2}$, $(N_i/n_{\rm obs})^{1/2}$	$C^{-1/3}$	R, $R_{\rm free}$	$d_{\min}$ (Å)	DPI $\sigma(r, B_{\rm avg})$ (Å)	Luzzati $\langle\Delta r\rangle$ (Å)	Read $\langle\Delta r\rangle$ (Å)	Reference
Concanavalin A	2130	116712	0.148 0.135	1.099	0.128 0.148	0.94	0.034 0.036	0.06		(*a*)
γB-Crystallin	1708	26151	0.297 0.256	1.032	0.180 0.204	1.49	0.14 0.14	0.16	0.12	(*b*)
βB2-Crystallin	1558	18583	0.356 0.290	~1.032	0.184 0.200	2.10	0.25 0.22	0.21	0.17	(*b*)
Ribonuclease A with RI	4416	18859	1.922 0.484	1.145	0.194 0.286	2.50	1.85 0.69	0.32	0.57	(*c*)
Fab HyHEL-5 with HEWL	4333	11754	† 0.607	1.111	0.196 0.288	2.65	— 0.69	0.30		(*d*)

References: (*a*) Deacon *et al.* (1997); (*b*) Tickle *et al.* (1998*a*); (*c*) Kobe & Deisenhofer (1995); (*d*) Cohen *et al.* (1996). † (N_i/p) negative.

r.m.s. position difference between the structures for atoms with $B_i = B_{\rm avg}$ of $(0.16^2 + 0.24^2)^{1/2} = 0.29$ Å. Daopin *et al.* reported the differences between the two determinations, omitting poor parts, as $\langle\Delta r\rangle_{\rm rms} = 0.15$ Å (main chain) and 0.29 Å (all atoms).

Human diferric lactoferrin (Haridas *et al.*, 1995) is an example of a large protein at the lower resolution of 2.2 Å, with a high value of $(N_i/p)^{1/2}$, leading to $\sigma(r, B_{\rm avg}) = 0.43$ Å.

Three crystal forms of thaumatin were studied by Ko *et al.* (1994). The orthorhombic and tetragonal forms diffracted to 1.75 Å, but the monoclinic $C2$ form diffracted only to 2.6 Å. The structures with 1552 protein atoms were successfully refined with restraints by *XPLOR* and *TNT*. For the monoclinic form, the number of parameters exceeds the number of diffraction observations, so (N_i/p) is negative and no estimate by (18.5.6.9) of the diffraction-data-only error is possible. The DPI (18.5.6.9) gives 0.17 and 0.16 Å for the orthorhombic and tetragonal forms, respectively.

18.5.7.3. Examples of the DPI using R_{free}

As in the case of monoclinic thaumatin, for low-resolution structures the number of parameters may exceed the number of diffraction data. To circumvent this difficulty, it was proposed in Section 18.5.6.3 to replace $p = n_{\rm obs} - n_{\rm params}$ by $n_{\rm obs}$ and R by $R_{\rm free}$ in a revised formula (18.5.6.10) for the DPI. Table 18.5.7.3 shows examples for some structures for which both R and $R_{\rm free}$ were available. The second row for each protein shows the alternative values for $(N_i/n_{\rm obs})^{1/2}$, $R_{\rm free}$ and the DPI $\sigma(r, B_{\rm avg})$ from (18.5.6.10).

For the structures with $d_{\min} \leq 2.0$ Å, the DPI is much the same whether it is based on R or $R_{\rm free}$.

Tickle *et al.* (1998*a*) have made full-matrix error estimates for isotropic restrained refinements of γB-crystallin with $d_{\min} = 1.49$ Å and of βB2-crystallin with $d_{\min} = 2.10$ Å. The DPI $\sigma(r, B_{\rm avg})$ calculated for the two structures is 0.14 and 0.25 Å with R in (18.5.6.9), and 0.14 and 0.22 Å with $R_{\rm free}$ in (18.5.6.10). The full-matrix weighted averages of $\sigma_{\rm res}(r)$ for all protein atoms were 0.10 and 0.15 Å, for only main-chain atoms 0.05 and 0.08 Å, for side-chain atoms 0.14 and 0.20 Å, and for water oxygens 0.27 and 0.35 Å. Again, the DPI gives reasonable overall indices for the quality of the structures.

For the complex of bovine ribonuclease A and porcine ribonuclease inhibitor (Kobe & Deisenhofer, 1995) with $d_{\min} =$ 2.50 Å, the number of reflections is only just larger than the number of parameters, so that $(N_i/p)^{1/2} = 1.922$ is very large, and the DPI with R gives an unrealistic 1.85 Å. With $R_{\rm free}$, $\sigma(r, B_{\rm avg}) = 0.69$ Å.

The HyHEL-5–lysozyme complex (Cohen *et al.*, 1996) had $d_{\min} = 2.65$ Å. Here the number of reflections is less than the number of parameters, but the $R_{\rm free}$ formula gives $\sigma(r, B_{\rm avg}) =$ 0.69 Å.

18.5.7.4. Comments on the diffraction-component precision index

The DPI (18.5.6.9) or (18.5.6.10) provides a very simple formula for $\sigma(r, B_{\rm avg})$. It is based on a very rough approximation to a diagonal element of the diffraction-data-only matrix. Using a diagonal element is a reasonable approximation for atomic resolution structures, but for low-resolution structures there will be significant off-diagonal terms between overlapping atoms. The effect can be simulated in the two-atom protein model of Section 18.5.3.2 by introducing positive off-diagonal elements into the diffraction-data matrix (18.5.3.3). As expected, $\sigma^2_{\rm diff}(x_i)$ is increased. So the DPI will be an underestimate of the diffraction component in low-resolution structures.

However, the true restrained variance $\sigma^2_{\rm res}(x_i)$ in the new counterpart of (18.5.3.12) remains less than the diagonal diffraction result (18.5.3.11) $\sigma^2_{\rm diff}(x_i) = 1/a$. Thus for low-resolution structures, the DPI should be an overestimate of the true precision given by a restrained full-matrix calculation (where the restraints act to hold the overlapping atoms apart). This is confirmed by the results for the 2.1 Å study of βB2-crystallin (Tickle *et al.*, 1998*a*) discussed in Section 18.5.7.3 and Table 18.5.7.3. The restrained full-matrix average for all protein atoms was $\sigma_{\rm res}(r) = 0.15$ Å, compared with the DPI 0.25 Å (on R) or 0.22 Å (on $R_{\rm free}$). The ratio between the unrestrained DPI and the restrained full-matrix average is consistent with a view of a low-resolution protein as a chain of effectively rigid peptide groups. The ratio no doubt gets much worse for resolutions of 3 Å and above.

The DPI estimate of $\sigma(r, B_{\rm avg})$ is given by a formula of 'back-of-an-envelope' simplicity. $B_{\rm avg}$ is taken to be the average B for fully occupied sites, but the weights implicit in the averaging are not well defined in the derivation of the DPI. Thus the DPI should perhaps be regarded as simply offering an estimate of a typical $\sigma_{\rm diff}(r)$ for a carbon or nitrogen atom with a mid-range B.

Table 18.5.8.1
$R = \langle|\Delta F|\rangle/\langle|F|\rangle$ as a function of $s\langle\Delta r\rangle$ in the Luzzati model for three-dimensional noncentrosymmetric structures ($s = 2\sin\theta/\lambda$)

$s\langle\Delta r\rangle$	R	$s\langle\Delta r\rangle$	R
0.00	0.000	0.10	0.237
0.01	0.025	0.12	0.281
0.02	0.050	0.14	0.319
0.03	0.074	0.16	0.353
0.04	0.098	0.18	0.385
0.05	0.122	0.20	0.414
0.06	0.145	0.25	0.474
0.07	0.168	0.30	0.518
0.08	0.191	0.35	0.548
0.09	0.214	∞	0.586

From the evidence of the tables in this section, except at low resolution, it seems to give a useful overall indication of protein precision, even in restrained refinements.

The DPI evidently provides a method for the comparative ranking of different structure determinations. In this regard it is a complement to the general use of $d_{\rm min}$ as a quick indicator of possible structural quality.

Note that (18.5.6.3) and (18.5.6.4) offer scope for making individual error estimates for atoms of different B and Z.

18.5.8. Luzzati plots

18.5.8.1. *Luzzati's theory*

Luzzati (1952) provided a theory for estimating, at any stage of a refinement, the average positional shifts which would be needed in an idealized refinement to reach $R = 0$. He did not provide a theory for estimating positional errors at the end of a normal refinement.

(1) His theory assumed that the $F_{\rm obs}$ had no errors, and that the $F_{\rm calc}$ model (scattering factors, thermal parameters *etc.*) was perfect, apart from coordinate errors.
(2) The Gaussian probability distribution for these coordinate errors was assumed to be the *same for all atoms*, independent of Z or B.
(3) The atoms were not required to be identical, and the position errors were not required to be small.

Luzzati gave families of curves for R *versus* $2\sin\theta/\lambda$ for varying average positional errors $\langle\Delta r\rangle$ for both centrosymmetric and noncentrosymmetric structures. The curves do not depend on the number N of atoms in the cell. They all rise from $R = 0$ at $2\sin\theta/\lambda = 0$ to the Wilson (1950) values 0.828 and 0.586 for random structures at high $2\sin\theta/\lambda$. Table 18.5.8.1 gives $R = \langle|\Delta F|\rangle/\langle|F|\rangle$ as a function of $s\langle\Delta r\rangle$ for three-dimensional noncentrosymmetric structures.

In a footnote (p. 807), Luzzati suggested that at the end of a normal refinement (with R nonzero due to experimental and model errors, *etc.*), the curves would indicate an upper limit for $\langle\Delta r\rangle$. He noted that typical small-molecule $\sigma(r)$'s of 0.01–0.02 Å, if used as $\langle\Delta r\rangle$ in the plots, would give much smaller R's than are found at the end of a refinement.

As examples, the Luzzati plots for the two structures of TGF-β2 are shown in Fig. 18.5.8.1. Daopin *et al.* (1994) inferred average $\langle\Delta r\rangle$'s around 0.21 Å for 1TGI and 0.23 Å for 1TGF.

Of the three Luzzati assumptions summarized above, the most attractive is the third, which does not require the atoms to be identical nor the position errors to be small. For proteins, there are very obvious difficulties with assumption (2). Errors do depend very strongly on Z and B. In the high-angle data shells, atoms with large B's contribute neither to ΔF nor to $|F|$, and so have no effect on R in these shells. In their important paper on protein accuracy, Chambers & Stroud (1979) said 'the [Luzzati] estimate derived from reflections in this range applies mainly to [the] best determined atoms.'

Thus a Luzzati plot seems to allow a cautious upper-limit statement about the precision of the best parts of a structure, but it gives little indication for the poor parts.

One reason for the past popularity of Luzzati plots has been that the R values for the middle and outer shells of a structure often roughly follow a Luzzati curve. Evidently, the effective average $\langle\Delta r\rangle$ for the structure must be decreasing as $2\sin\theta/\lambda$ increases, since atoms of high B are ceasing to contribute, whereas the proportionate experimental errors must be increasing. This also suggests that the upper limit for $\langle\Delta r\rangle$ for the low-B atoms could be estimated from the lowest Luzzati theoretical curve touched by the experimental R plot. Thus in Fig. 18.5.8.1 the upper limits for the low-B atoms could be taken as 0.18 and 0.21 Å, rather than the 0.21 and 0.23 Å chosen by Daopin *et al.*

From the introduction of $R_{\rm free}$ by Brünger (1992) and the discussion of $R_{\rm free}$ by Tickle *et al.* (1998*b*), it can be seen that Luzzati plots should be based on a residual more akin to $R_{\rm free}$ than R in order to avoid bias from the fitting of data.

The mean positional error $\langle\Delta r\rangle$ of atoms can also be estimated from the σ_A plots of Read (1986, 1990). This method arose from Read's analysis of improved Fourier coefficients for maps using phases from partial structures with errors. It is preferable in several respects to the Luzzati method, but like the Luzzati method it assumes that the coordinate distribution is the same for all atoms. Luzzati and/or Read estimates of $\langle\Delta r\rangle$ are available for some of the structures in Tables 18.5.7.2 and 18.5.7.3. Often, the two estimates are not greatly different.

18.5.8.2. *Statistical reinterpretation of Luzzati plots*

Luzzati plots are fundamentally different from other statistical estimates of error. The Luzzati theory applies to an idealized incomplete refinement and estimates the average shifts needed to reach $R = 0$. In the least-squares method, the equations for shifts are quite different from the equations for estimating variances in a converged refinement. However, Luzzati-style plots of R *versus* $2\sin\theta/\lambda$ can be reinterpreted to give statistically based estimates of $\sigma(x)$.

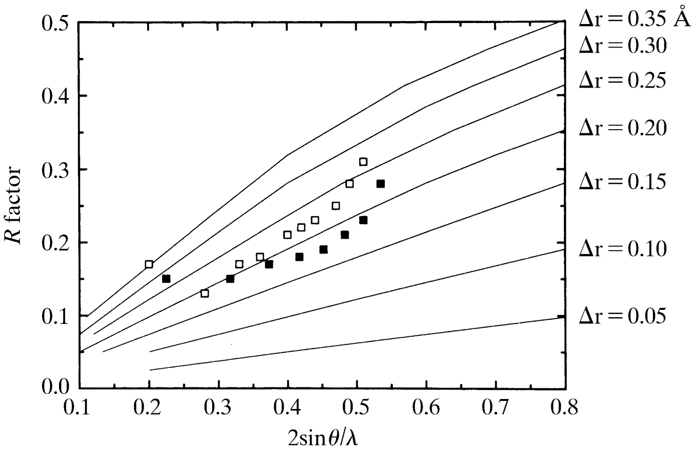


Figure 18.5.8.1
Luzzati plots showing the refined R factor as a function of resolution for 1TGI (solid squares) and 1TGF (open squares) (Daopin *et al.*, 1994).

During Cruickshank's (1960) derivation of the approximate equation (18.5.6.2) for $\sigma(x)$ in diagonal least squares, he reached an intermediate equation

$$\sigma^2(x) = N_i \bigg/ \left[4 \sum_{\rm obs} (s^2/R^2)\right]. \qquad (18.5.8.1)$$

He then assumed R to be independent of s $(= 2\sin\theta/\lambda)$ and took R outside the summation to reach (18.5.6.2) above.

Luzzati (1952) calculated the acentric residual R as a function of $\langle\Delta r\rangle$, the average radial error of the atomic positions. His analysis shows that R is a linear function of s and $\langle\Delta r\rangle$ for a substantial range of $s\langle\Delta r\rangle$, with

$$R(s,\ \langle\Delta r\rangle) = (2\pi)^{1/2} s\langle\Delta r\rangle. \qquad (18.5.8.2)$$

The theoretical Luzzati plots of R are nearly linear for small-to-medium $s = 2\sin\theta/\lambda$ (see Fig. 18.5.8.1). If we substitute this R in the least-squares estimate (18.5.8.1) and use the three-dimensional-Gaussian relation $\sigma(r) = 1.085\langle\Delta r\rangle$, some manipulation (Cruickshank, 1999) along the lines of Section 18.5.6 eventually yields a statistically based formula,

$$\sigma_{\rm LS,Luzz}(r) = 1.33(N_i/p)^{1/2}[R(s_m)/s_m], \qquad (18.5.8.3)$$

where $R(s_m)$ is the value of R at some value of $s = s_m$ on the selected Luzzati curve. Equation (18.5.8.3) provides a means of making a very rough statistical estimate of error for an atom with $B = B_{\rm avg}$ (the average B for fully occupied sites) from a plot of R *versus* $2\sin\theta/\lambda$.

The corresponding equation involving $R_{\rm free}$ is

$$\sigma_{\rm LS,Luzz}(r) = 1.33(N_i/n_{\rm obs})^{1/2}[R_{\rm free}(s_m)/s_m]. \qquad (18.5.8.4)$$

18.5.8.3. Comments on Luzzati plots

Protein structures always show a great range of B values. The Luzzati theory effectively assumes that all atoms have the same B. Nonetheless, the Luzzati method applied to high-angle data shells does provide an upper limit for $\langle\Delta r\rangle$ for the atoms with low B. It is an upper limit since experimental errors and model imperfections are not allowed for in the theory.

Low-resolution structures can be determined validly by using restraints, even though the number of diffraction observations is less than the number of atomic coordinates. The Luzzati method, based preferably on $R_{\rm free}$, can be applied to the atoms of low B in such structures. As the number of observations increases, and the resolution improves, the Luzzati $\langle\Delta r\rangle$ increasingly overestimates the true $\sigma(r)$ of the low-B atoms.

In the use of Luzzati plots, the method of refinement, and its degree of convergence, is irrelevant. A Luzzati plot is a statement for the low-B atoms about the maximum errors associated with a given structure, whether converged or not.

References

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*a*). *A systematic study of coordinate precision in X-ray structure analyses. I. Descriptive statistics and predictive estimates of e.s.d.'s for C atoms. Acta Cryst.* A**51**, 95–111.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*b*). *A systematic study of coordinate precision in X-ray structure analyses. II. Predictive estimates of e.s.d.'s for the general-atom case. Acta Cryst.* A**51**, 112–121.

Bricogne, G. (1993). *Direct phase determination by entropy maximization and likelihood ranking: status report and perspectives. Acta Cryst.* D**49**, 37–60.

Bricogne, G. (2008). *Fourier transforms in crystallography: theory, algorithms, and applications*. In *International Tables for Crystallography*, Vol. B, edited by U. Shmueli, ch. 1.3. Heidelberg: Springer.

Bricogne, G. & Irwin, J. (1996). *Maximum-likelihood structure refinement: theory and implementation within BUSTER + TNT*. In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *Free R-value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and application. Acta Cryst.* D**49**, 24–36.

Chambers, J. L. & Stroud, R. M. (1979). *The accuracy of refined protein structures: comparison of two independently refined models of bovine trypsin. Acta Cryst.* B**35**, 1861–1874.

Cochran, W. (1948). *The Fourier method of crystal-structure analysis. Acta Cryst.* **1**, 138–142.

Cohen, G. H., Sheriff, S. & Davies, D. R. (1996). *Refined structure of the monoclonal antibody HyHEL-5 with its antigen hen egg-white lysozyme. Acta Cryst.* D**52**, 315–326.

Cowtan, K. & Ten Eyck, L. F. (2000). *Eigensystem analysis of the refinement of a small metalloprotein. Acta Cryst.* D**56**, 842–856.

Cruickshank, D. W. J. (1949*a*). *The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1949*b*). *The accuracy of atomic coordinates derived by least-squares or Fourier methods. Acta Cryst.* **2**, 154–157.

Cruickshank, D. W. J. (1952). *On the relations between Fourier and least-squares methods of structure determination. Acta Cryst.* **5**, 511–518.

Cruickshank, D. W. J. (1956). *The determination of the anisotropic thermal motion of atoms in crystals. Acta Cryst.* **9**, 747–753.

Cruickshank, D. W. J. (1959). *Statistics*. In *International Tables for X-ray Crystallography*, Vol. 2, edited by J. S. Kasper & K. Lonsdale, pp. 84–98. Birmingham: Kynoch Press.

Cruickshank, D. W. J. (1960). *The required precision of intensity measurements for single-crystal analysis. Acta Cryst.* **13**, 774–777.

Cruickshank, D. W. J. (1965). *Notes for authors; anisotropic parameters. Acta Cryst.* **19**, 153.

Cruickshank, D. W. J. (1970). *Least-squares refinement of atomic parameters*. In *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 187–196. Copenhagen: Munksgaard.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Cruickshank, D. W. J. & Robertson, A. P. (1953). *The comparison of theoretical and experimental determinations of molecular structures, with applications to naphthalene and anthracene. Acta Cryst.* **6**, 698–705.

Cruickshank, D. W. J. & Rollett, J. S. (1953). *Electron-density errors at special positions. Acta Cryst.* **6**, 705–707.

Daopin, S., Davies, D. R., Schlunegger, M. P. & Grütter, M. G. (1994). *Comparison of two crystal structures of TGF-β2: the accuracy of refined protein structures. Acta Cryst.* D**50**, 85–92.

Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *The structure of concanavalin A and its bound solvent determined with small-molecule accuracy at 0.94 Å resolution. J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.

Dodson, E. (1998). *The role of validation in macromolecular crystallography. Acta Cryst.* D**54**, 1109–1118.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Haridas, M., Anderson, B. F. & Baker, E. N. (1995). *Structure of human diferric lactoferrin refined at 2.2 Å resolution. Acta Cryst.* D**51**, 629–646.

Hendrickson, W. A. & Konnert, J. H. (1980). *Incorporation of stereochemical information into crystallographic refinement*. In *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.23. Bangalore: Indian Academy of Sciences.

International Tables for Crystallography (2004). Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince. Dordrecht: Kluwer Academic Publishers.

Jiang, J.-S. & Brünger, A. T. (1994). *Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures. J. Mol. Biol.* **243**, 100–115.

Ko, T.-P., Day, J., Greenwood, A. & McPherson, A. (1994). *Structures of three crystal forms of the sweet protein thaumatin. Acta Cryst.* D**50**, 813–825.

Kobe, B. & Deisenhofer, J. (1995). *A structural basis of the interactions between leucine-rich repeats and protein ligands. Nature (London)*, **374**, 183–186.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *The molecular configuration of deoxyribonucleic acid. II. Molecular models and their Fourier transforms. J. Mol. Biol.* **2**, 38–64.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures cristallines. Acta Cryst.* **5**, 802–810.

Moews, P. C. & Kretsinger, R. H. (1995). *Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. J. Mol. Biol.* **91**, 201–228.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Efficient anisotropic refinement of macromolecular structures using FFT. Acta Cryst.* D**55**, 247–255.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.

Rollett, J. S. (1970). *Least-squares procedures in crystal structure analysis*. In *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 167–181. Copenhagen: Munksgaard.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Gonschorek, W., Hahn, Th., Huml, K., Marsh, R. E., Prince, E., Robertson, B. E., Rollett, J. S. & Wilson, A. J. C. (1989). *Statistical descriptors in crystallography: Report of the IUCr subcommittee on statistical descriptors. Acta Cryst.* A**45**, 63–75.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Statistical descriptors in crystallography. II. Report of a working group on expression of uncertainty in measurement. Acta Cryst.* A**51**, 565–569.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Ribonuclease from Streptomyces aureofaciens at atomic resolution. Acta Cryst.* D**52**, 327–344.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high resolution refinement. Methods Enzymol.* **277**, 319–343.

Stec, B., Zhou, R. & Teeter, M. M. (1995). *Full-matrix refinement of the protein crambin at 0.83 Å and 130 K. Acta Cryst.* D**51**, 663–681.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*a*). *Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γB- and βB2-crystallin. Acta Cryst.* D**54**, 243–252.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*b*). *R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.

Tronrud, D. E. (1997). *TNT refinement package. Methods Enzymol.* **277**, 306–319.

Tronrud, D. E. (1999). *The efficient calculation of the normal matrix in least-squares refinement of macromolecular structures. Acta Cryst.* A**55**, 700–703.

Trueblood, K. N., Bürgi, H.-B., Burzlaff, H., Dunitz, J. D., Gramaccioli, C. M., Schulz, H. H., Shmueli, U. & Abrahams, S. C. (1996). *Atomic displacement parameter nomenclature. Report of a subcommittee on atomic displacement parameter nomenclature. Acta Cryst.* A**52**, 770–781.

Usón, I., Pohl, E., Schneider, T. R., Dauter, Z., Schmidt, A., Fritz, H.-J. & Sheldrick, G. M. (1999). *1.7 Å structure of the stabilized REI_V mutant T39K. Application of local NCS restraints. Acta Cryst.* D**55**, 1158–1167.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Refinement of the model of a protein: rubredoxin at 1.5 Å resolution. Acta Cryst.* B**29**, 943–956.

Wilson, A. J. C. (1950). *Largest likely values for the reliability index. Acta Cryst.* **3**, 397–398.

International Tables for Crystallography (2012). Vol. F, Chapter 18.6, pp. 512–519.

Chapter 18.6. *CNS*, a program system for structure-determination and refinement

A. T. Brunger, P. D. Adams, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, N. S. Pannu, R. J. Read, L. M. Rice and T. Simonson

18.6.1. Introduction

We have developed a new and advanced software system, the *Crystallography & NMR System* (CNS), for crystallographic and NMR structure determination (Brünger *et al.*, 1998). The goals of CNS are: (1) to create a flexible computational framework for exploration of new approaches to structure determination; (2) to provide tools for structure solution of difficult or large structures; (3) to develop models for analysing structural and dynamical properties of macromolecules; and (4) to integrate all sources of information into all stages of the structure-determination process.

To meet these goals, algorithms were moved from the source code into a symbolic structure-determination language which represents a new concept in computational crystallography. The high-level CNS computing language allows definition of symbolic target functions, data structures, procedures and modules. The CNS program acts as an interpreter for the high-level CNS language and includes hard-wired functions for efficient processing of computing-intensive tasks. Methods and algorithms are therefore more clearly defined and easier to adapt to new and challenging problems. The result is a multi-level system which provides maximum flexibility to the user (Fig. 18.6.1.1). The CNS language provides a common framework for nearly all computational procedures of structure determination. A comprehensive set of crystallographic procedures for phasing, density modification and refinement has been implemented in this language. Task-oriented input files written in the CNS language, which can also be accessed through an HTML graphical interface (Graham, 1995), are available to carry out these procedures.

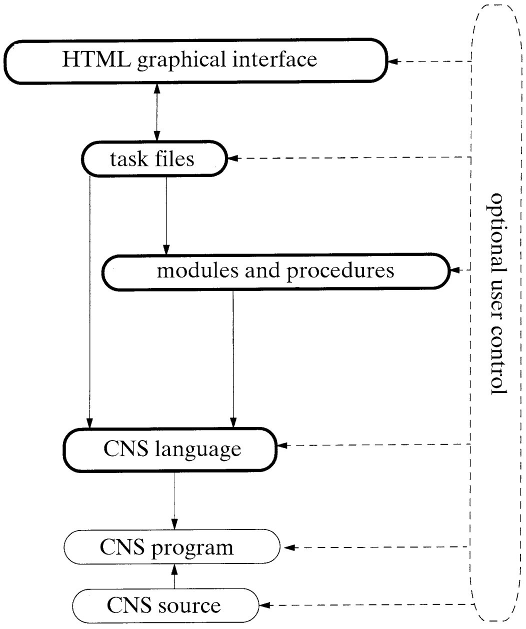


Figure 18.6.1.1
CNS consists of five layers which are under user control. The high-level HTML graphical interface interacts with the task-oriented input files. The task files use the CNS language and the modules. The modules contain CNS language statements. The CNS language is interpreted by the CNS Fortran77 program. The program performs the data manipulations, data operations and 'hard-wired' algorithms.

18.6.2. The CNS language

One of the key features of the CNS language is symbolic data structure manipulation, for example,

```
xray
  do (pa = −2 * (amplitude(fp)^2 + amplitude(fh)^2
           − amplitude(fph) ^2) * amplitude(fp)
           * real(fh)/(3 * v^2 + 4 * (amplitude(fph)^2
           + sph^2) * v)) (acentric)
end
```

(18.6.2.1)

which is equivalent to the following mathematical expression for all acentric indices **h**,

$$p_a(\mathbf{h}) = 2\frac{-[|\mathbf{f}_p(\mathbf{h})|^2 + |\mathbf{f}_h(\mathbf{h})|^2 - |\mathbf{f}_{ph}(\mathbf{h})|^2]|\mathbf{f}_p(\mathbf{h})|\{[\mathbf{f}_h(\mathbf{h}) + \mathbf{f}_h(\mathbf{h})^*]/2\}}{3v(\mathbf{h})^2 + 4[|\mathbf{f}_{ph}(\mathbf{h})|^2 + s_{ph}(\mathbf{h})^2]v(\mathbf{h})}, \tag{18.6.2.2}$$

where $\mathbf{f}_p$ ['fp' in equation (18.6.2.1)] is the 'native' structure-factor array, $\mathbf{f}_{ph}$ ['fph' in equation (18.6.2.1)] is the derivative structure-factor array, s_{ph} ['sph' in equation (18.6.2.1)] is the corresponding experimental σ, v is the expectation value for the lack of closure (including lack of isomorphism and errors in the heavy-atom model), and $\mathbf{f}_h$ ['fh' in equation (18.6.2.1)] is the calculated heavy-atom structure-factor array. This expression computes the $A_{\rm iso}$ coefficient of the phase probability distribution for single isomorphous replacement described by Hendrickson & Lattman (1970) and Blundell & Johnson (1976).

The expression in equation (18.6.2.1) is computed for the specified subset of reflections '(acentric)'. This expression means that only the selected (in this case all acentric) reflections are used. More sophisticated selections are possible, *e.g.*

```
(amplitude(fp)>2 * sh and amplitude(fph)>2 * sph
  and d> = 3)
```

(18.6.2.3)

selects all reflections with Bragg spacing, d, greater than 3 Å for which both native (fp) and derivative (fph) amplitudes are greater than two times their corresponding σ values ('sh' and 'sph', respectively). Extensive use of this structure-factor selection facility is made for cross-validating statistical properties, such as R values (Brünger, 1992), σ_A values (Kleywegt & Brünger, 1996; Read, 1997) and maximum-likelihood functions (Pannu & Read, 1996; Adams *et al.*, 1997).

Similar operations exist for electron-density maps, *e.g.*

```
xray
  do (map = 0) (map < 0.1)
end
```

(18.6.2.4)

is an example of a truncation operation: all map values less than

```
evaluate ( $crystal_lattice.space_group     = "P2(1)2(1)2(1)" )
evaluate ( $crystal_lattice.unit_cell.a     = 61.76           )
evaluate ( $crystal_lattice.unit_cell.b     = 40.73           )
evaluate ( $crystal_lattice.unit_cell.c     = 26.74           )
evaluate ( $crystal_lattice.unit_cell.alpha = 90              )
evaluate ( $crystal_lattice.unit_cell.beta  = 90              )
evaluate ( $crystal_lattice.unit_cell.gamma = 90              )
```

(*a*)

```
define (
   &crystal_lattice.space_group     = P2(1)2(1)2(1) ;
   &crystal_lattice.unit_cell.a     = 61.76 ;
   &crystal_lattice.unit_cell.b     = 40.73 ;
   &crystal_lattice.unit_cell.c     = 26.74 ;
   &crystal_lattice.unit_cell.alpha = 90 ;
   &crystal_lattice.unit_cell.beta  = 90 ;
   &crystal_lattice.unit_cell.gamma = 90 ;
)
```

(*b*)

Figure 18.6.3.1
Examples of compound symbols and compound parameters. (*a*) The ‘evaluate’ statement is used to define typed symbols (strings, numbers and logicals). Symbol names are in bold. (*b*) The ‘define’ statement is used to define untyped parameters. Each parameter entry is terminated by a semicolon. The compound base name ‘crystal_lattice’ has a number of sub-levels, such as ‘space_group’ and the ‘unit_cell’ parameters. ‘unit_cell’ is itself base to a number of sub-levels, such as ‘a’ and ‘alpha’. Parameter names are in bold.

```
do (eobs=norm(amplitude(fobs)))                              ( test_set )
do (ecalc=norm(amplitude(fcalc)))                            ( test_set )
do (sigmaA=sigacv(eobs,ecalc))                               ( test_set )
do (sigmaD=sqrt(save(amplitude(fobs))^2 (1-sigmaA^2)))       ( test_set )
do (D= sigmaA * sqrt(save(amplitude(fobs))^2  /
            save(amplitude(fcalc)^2)))                       ( test_set )

do (sigmaA=sum(sigmaA*test)/ max(1,sum(test)))               ( all )
do (sigmaD=sum(sigmaD*test)/ max(1,sum(test)))               ( all )
do (D=sum(D*test)/ max(1,sum(test)))                         ( all )
```

Figure 18.6.4.1
Example for statistical operations provided by the CNS language. ‘norm’, ‘sigacv’, ‘save’ and ‘sum’ are functions that are computed internally by the CNS program. Binwise operations are in italics (‘sigacv’, ‘save’ and ‘sum’). The result for a particular bin is stored in all elements belonging to the bin. The σ_A (‘sigmaA’) parameters are computed in binwise resolution shells. The σ_Δ (‘sigmaD’) and D parameters are then computed from σ_A and binwise averages involving $|\mathbf{F}_o|^2$ and $|\mathbf{F}_c|^2$. The binwise results are expanded to all reflections by the last three statements. ‘test’ is an array that is 1 for all reflections in the test set and 0 otherwise. ‘sum’ is a binwise operation on all reflections with the same partitioning used for the test set.

0.1 are set to 0. Atoms can be selected based on a number of atomic properties and descriptors, *e.g.*

```
do (b = 10)(residue 1:40 and
            (name ca or name n or name c or name o))
```
(18.6.2.5)

sets the B factors of all polypeptide backbone atoms of residues 1 through 40 to 10 Å^2.

Operations exist between data structures, *e.g.* real- and reciprocal-space arrays, and atom properties. For example, Fourier transformations between real and reciprocal space can be accomplished by the following CNS commands:

```
xray
  mapresolution infinity 3.
  fft grid 0.3333 end
  do (map = ft(f_cal)) (acentric)
end
```
(18.6.2.6)

which computes a map on a 1 Å grid by Fourier transformation of the ‘f_cal’ array for all acentric reflections.

Atoms can be associated with calculated structure factors, *e.g.*

```
associate f_cal (residue 1:50)
```
(18.6.2.7)

This statement will associate the reciprocal-space array ‘f_cal’ with the atoms belonging to residues 1 through 50. These structure-factor associations are used in the symbolic target functions described below.

There are no predefined reciprocal- or real-space arrays in CNS. Dynamic memory allocation allows one to carry out operations on arbitrarily large data sets with many individual entries (*e.g.* derivative diffraction data) without the need to recompile the source code. The various reciprocal-space structure-factor arrays must therefore be declared and their type specified prior to invoking them. For example, a reciprocal-space array with real values, such as observed amplitudes, is declared by

```
declare name = fobs type = real domain = reciprocal end
```
(18.6.2.8)

Reciprocal-space arrays can be grouped. For example, Hendrickson & Lattman (1970) coefficients are represented as a group of four reciprocal-space structure-factor arrays,

```
group type = hl object = pa object = pb
  object = pc object = pd end
```
(18.6.2.9)

where ‘pa’, ‘pb’, ‘pc’ and ‘pd’ refer to the individual arrays. This group statement indicates to CNS that the specified arrays need to be transformed together when reflection indices are changed, *e.g.* during expansion of the diffraction data to space group *P*1.

18.6.3. Symbols and parameters

The CNS language supports two types of data elements which may be used to store and retrieve information. *Symbols* are typed variables, such as numbers, character strings of restricted length and logicals. *Parameters* are untyped data elements of arbitrary length that may contain collections of CNS commands, numbers, strings or symbols.

Symbols are denoted by a dollar sign ($), and parameters by an ampersand (&). Symbols and parameters may contain a single data element, or they may be a *compound* data structure of arbitrary complexity. The hierarchy of these data structures is denoted using a period (.). Figs. 18.6.3.1(*a*) and (*b*) demonstrate how crystal-lattice information can be stored in compound symbols and parameters, respectively. The information stored in symbols or parameters can be retrieved by simply referring to them within a CNS command: the symbol or parameter name is substituted by its content. Symbol substitution of portions of the compound names (*e.g.* ‘&crystal_lattice.unit_cell.$para’) allows one to carry out conditional and iterative operations on data structures, such as matrix multiplication.

18.6.4. Statistical functions

The CNS language contains a number of statistical operations, such as binwise averages and summations. The resolution bins are defined by a central facility in CNS.

Fig. 18.6.4.1 shows how σ_A, σ_Δ and D (Read, 1986, 1990) are computed from the observed structure factors (‘fobs’) and the calculated model structure factors (‘fcalc’) using the CNS statistical operations. The first five operations are performed for the reflections in the test set, while the last three operations expand the results to all reflections. The ‘norm’ function computes normalized structure-factor amplitudes for the specified argu-

```
associate f_h_1 <atom-selection-1>
associate f_h_2 <atom-selection-2>
associate f_h_3 <atom-selection-3>

target=(
         (abs(f_h_1+f_p)-f_ph_1)^2 / (2*v_1) +
         (abs(f_h_2+f_p)-f_ph_2)^2 / (2*v_2) +
         (abs(f_h_3+f_p)-f_ph_3)^2 / (2*v_3)
       )

dtarget(f_h_1)=
        (
         2*(abs(f_h_1+f_p)-f_ph_1)
           *(f_h_1+f_p)/abs(f_h_1+f_p) / (2*v_1)
         )

dtarget(f_h_2)=
        (
         2*(abs(f_h_2+f_p)-f_ph_2)
           *(f_h_2+f_p)/abs(f_h_2+f_p) / (2*v_2)
         )

dtarget(f_h_3)=
        (
         2*(abs(f_h_3+f_p)-f_ph_3)
           *(f_h_3+f_p)/abs(f_h_3+f_p) / (2*v_3)

tselection=<selection>
cvselection=<selection>
```

(*a*)

```
associate fcalc1 <atom-selection1>
associate fcalc2 <atom-selection2>

target=( abs(fobs) - sqrt(abs(fcalc1)^2+abs(fcalc2)^2))^2 )

dtarget(fcalc1)=( -2 *
            (abs(fobs)-sqrt(abs(fcalc1)^2+abs(fcalc2)^2)) *
              (fcalc1/(sqrt(abs(fcalc1)^2+abs(fcalc2)^2))))

dtarget(fcalc2)=( -2 *
            (abs(fobs)-sqrt(abs(fcalc1)^2+abs(fcalc2)^2)) *
              (fcalc2/(sqrt(abs(fcalc1)^2+abs(fcalc2)^2))))

tselection=<selection>
cvselection=<selection>
```

(*b*)

Figure 18.6.5.1
Examples of symbolic definition of a refinement target function and its derivatives with respect to the calculated structure-factor arrays. (*a*) Simultaneous refinement of heavy-atom sites of three derivatives. The target function is defined by the 'target' expression. 'f_h_1', 'f_h_2' and 'f_h_3' (in bold) are complex structure factors corresponding to three sets of heavy atoms that are specified using atom selections [equation (18.6.2.7)]. The target function and its derivatives with respect to the three structure-factor arrays are defined symbolically using the structure-factor amplitudes of the native crystal, 'f_p', those of the derivatives, 'f_ph_1', 'f_ph_2', 'f_ph_3', the complex structure factors of the heavy-atom models, 'f_h_1', 'f_h_2', 'f_h_3', and the corresponding lack-of-closure variances, 'v_1', 'v_2' and 'v_3'. The summation over the selected stucture factors ('tselection') is performed implicitly. (*b*) Refinement of two independent models against perfectly twinned data. 'fcalc1' and 'fcalc2' are complex structure factors for the models that are related by a twinning operation (in bold). The target function and its derivatives with respect to the two structure-factor arrays are explicitly defined.

ments. The 'sigacv' function evaluates σ_A from the normalized structure factors. The 'save' function computes the statistical average

$$\text{save}(f) = \frac{\sum_{hkl} f_{hkl}(w/\varepsilon)}{\sum_{hkl} w}, \qquad (18.6.4.1)$$

where w is 1 and 2 for centric and acentric reflections, respectively, and ε is the statistical weight. The averages are computed binwise, and the result for a particular bin is stored in all selected reflections belonging to the bin.

18.6.5. Symbolic target function

One of the key innovative features of CNS is the ability to symbolically define target functions and their first derivatives for crystallographic searches and refinement. This allows one conveniently to implement new crystallographic methodologies as they are being developed.

The power of symbolic target functions is illustrated by two examples. In the first example, a target function is defined for simultaneous heavy-atom parameter refinement of three derivatives. The sites for each of the three derivatives can be disjoint or identical, depending on the particular situation. For simplicity, the Blow & Crick (1959) approach is used, although maximum-likelihood targets are also possible (see below). The heavy-atom sites are refined against the target

$$\sum_{hkl} \frac{(|\mathbf{F}_{h_1} + \mathbf{F}_p| - |\mathbf{F}_{ph_1}|)^2}{2v_1} + \frac{(|\mathbf{F}_{h_2} + \mathbf{F}_p| - |\mathbf{F}_{ph_2}|)^2}{2v_2} + \frac{(|\mathbf{F}_{h_3} + \mathbf{F}_p| - |\mathbf{F}_{ph_3}|)^2}{2v_3}. \qquad (18.6.5.1)$$

$\mathbf{F}_{h_1}$, $\mathbf{F}_{h_2}$ and $\mathbf{F}_{h_3}$ are complex structure factors corresponding to the three sets of heavy-atom sites, $\mathbf{F}_p$ represents the structure factors of the native crystal, $|\mathbf{F}_{ph_1}|$, $|\mathbf{F}_{ph_2}|$ and $|\mathbf{F}_{ph_3}|$ are the structure-factor amplitudes of the derivatives, and v_1, v_2 and v_3 are the variances of the three lack-of-closure expressions. The corresponding target expression and its first derivatives with respect to the calculated structure factors are shown in Fig. 18.6.5.1(*a*). The derivatives of the target function with respect to each of the three associated structure-factor arrays are specified with the 'dtarget' expressions. The 'tselection' statement specifies the selected subset of reflections to be used in the target function (*e.g.* excluding outliers), and the 'cvselection' statement specifies a subset of reflections to be used for cross-validation (Brünger, 1992) (*i.e.* the subset is not used during refinement but only as a monitor for the progress of refinement).

The second example is the refinement of a perfectly twinned crystal with overlapping reflections from two independent crystal lattices. Refinement of the model is carried out against the residual

$$\sum_{hkl} |\mathbf{F}_{\text{obs}}| - (|\mathbf{F}_{\text{calc1}}|^2 + |\mathbf{F}_{\text{calc2}}|^2)^{1/2}. \qquad (18.6.5.2)$$

The symbolic definition of this target is shown in Fig. 18.6.5.1(*b*). The twinning operation itself is imposed as a relationship between the two sets of selected atoms (not shown). This example assumes that the two calculated structure-factor arrays ('fcalc1' and 'fcalc2') that correspond to the two lattices have been appropriately scaled with respect to the observed structure factors, and the twinning fractions have been incorporated into the scale factors. However, a more sophisticated target function could be defined which incorporates scaling.

A major advantage of the symbolic definition of the target function and its derivatives is that any arbitrary function of structure-factor arrays can be used. This means that the scope of possible targets is not limited to least-squares targets. Symbolic definition of numerical integration over unknown variables (such as phase angles) is also possible. Thus, even complicated maximum-likelihood target functions (Bricogne, 1984; Otwinowski, 1991; Pannu & Read, 1996; Pannu *et al.*, 1998) can be defined using the CNS language. This is particularly valuable at the prototype stage. For greater efficiency, the standard maximum-likelihood targets are provided through CNS source code which can be accessed as functions in the CNS language. For example, the maximum-likelihood target function MLF (Pannu & Read, 1996) and its derivative with respect to the calculated structure factors are defined as

```
module { compute_unit_cell_volume }
(
   &cell;
   &volume;
)


evaluate ( $cabg.1=cos(&cell.alpha) )
evaluate ( $sabg.1=sin(&cell.alpha) )

evaluate ( $cabg.2=cos(&cell.beta) )
evaluate ( $sabg.2=sin(&cell.beta) )

evaluate ( $cabg.3=cos(&cell.gamma) )
evaluate ( $sabg.3=sin(&cell.gamma) )

evaluate ( &volume=&cell.a * &cell.b * &cell.c *
                   sqrt(1+2*$cabg.1*$cabg.2*$cabg.3
                        -$cabg.1^2-$cabg.2^2-$cabg.3^2) )
```

Figure 18.6.6.1
Use of compound parameters within a module. This module computes the unit-cell volume (Stout & Jensen, 1989) from the unit-cell geometry. Input and output parameter base names are in bold. Local symbols, such as cabg.1, are defined through 'evaluate' statements. The result is stored in the parameter '&volume' which is passed to the invoking task file or module.

$$\begin{aligned}\texttt{target} &= \texttt{(mlf (fobs,sigma,(fcalc + fbulk),}\\ &\qquad \texttt{d,sigma\_delta))}\\ \texttt{dtarget} &= \texttt{(dmlf (fobs,sigma,(fcalc + fbulk),}\\ &\qquad \texttt{d,sigma\_delta))}\end{aligned}\tag{18.6.5.3}$$

where 'mlf( )' and 'dmlf( )' refer to internal maximum-likelihood functions, 'fobs' and 'sigma' are the observed structure-factor amplitudes and corresponding σ values, 'fcalc' is the (complex) calculated structure-factor array, 'fbulk' is the structure-factor array for a bulk solvent model, and 'd' and 'sigma_delta' are the cross-validated D and σ_Δ functions (Read, 1990; Kleywegt & Brünger, 1996; Read, 1997) which are precomputed prior to invoking the MLF target function using the test set of reflections. The availability of internal Fortran subroutines for the most computing-intensive target functions and the symbolic definitions involving structure-factor arrays allow for maximal flexibility and efficiency. Other examples of available maximum-likelihood target functions include MLI (intensity-based maximum-likelihood refinement), MLHL [crystallographic model refinement with prior phase information (Pannu *et al.*, 1998)], and maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement (Otwinowski, 1991) and MAD phasing (Hendrickson, 1991; Burling *et al.*, 1996). Work is in progress to define target functions that include correlations between different heavy-atom derivatives (Read, 1994).

18.6.6. Modules and procedures

Modules exist as separate files and contain collections of CNS commands related to a particular task. In contrast, *procedures* can be defined and invoked from within any file. Modules and procedures share a similar parameter-passing mechanism for both input and output. Modules and procedures make it possible to write programs in the CNS language in a manner similar to that of a computing language, such as Fortran or C. CNS modules and procedures have defined sets of input (and output) parameters that are passed into them (or returned) when they are invoked. This enables long collections of CNS language statements to be broken down into modules for greater clarity of the underlying algorithm.

Parameters passed into a module or procedure inherit the scope of the calling task file or module, and thus they exhibit a behaviour analogous to most computing languages. Symbols defined within a module or procedure are purely local variables.

```
module { phase_distribution }
     (
   &fp;  {input: native data}
   &sp;  {input: native sigma}
   &sel; {input: selection of structure factors}
   &fh;  {input: name of heavy atom structure factors}
   &fph; {input: name of derivative data array}
   &sph; {input: name of derivative's sigma array}
   &var; {input: lack-of-isomorphism plus measurement errors}
   &pa;  {output: Hendrickson and Lattman A array}
   &pb;  {output: Hendrickson and Lattman B array}
   &pc;  {output: Hendrickson and Lattman C array}
   &pd;  {output: Hendrickson and Lattman D array}
     )

do (&pa= (cos(centric_phase)*(
        -(abs(&fh+combine(abs(&fp),centric_phase))-abs(&fph))^2/(2*&var)
        +(abs(&fh-combine(abs(&fp),centric_phase))-abs(&fph))^2/(2*&var)))
  ( centric and &sel )

do (&pb=(sin(centric_phase)*(
        -(abs(&fh+combine(abs(&fp),centric_phase))-abs(&fph))^2/(2*&var)
        +(abs(&fh-combine(abs(&fp),centric_phase))-abs(&fph))^2/(2*&var)))
  ( centric and &sel )

do (&pc=0) (centric and &sel)

do (&pd=0) (centric and &sel)

do (&pa=-2 * (amplitude(&fp)^2 + amplitude(&fh)^2 - amplitude(&fph)^2 )
          * amplitude(&fp) * real(&fh)/
          (3 * &var^2 + 4 * (amplitude(&fph)^2+&sph^2) * &var))
  ( acentric and &sel )

do (&pb=-2 * (amplitude(&fp)^2 + amplitude(&fh)^2 - amplitude(&fph)^2 )
          * amplitude(&fp) * imag(&fh)/
          (3 * &var^2 + 4 * (amplitude(&fph)^2+&sph^2) * &var) )
  ( acentric and &sel )

do (&pc=-amplitude(&fp)^2 * (real(&fh)^2 - imag(&fh)^2) /
        (3 * &var^2 + 4 * (amplitude(&fph)^2+&sph^2) * &var) )
  ( acentric and &sel )

do (&pd=-2 * amplitude(&fp)^2 * real(&fh) * imag(&fh) /
       (3 * &var^2 + 4 * (amplitude(&fph)^2+&sph^2) * &var) )
  ( acentric and &sel )
```

(*a*)

```
@phase_distribution
  (
   &fp=fobs;
   &sp=sigma;
   &sel=( d > 3. );
   &fh=f_heavy;
   &fph=f_deriv;
   &sph=s_deriv;
   &var=variance;
   &pa=pa;
   &pb=pb;
   &pc=pc;
   &pd=pd;
  )
```

(*b*)

Figure 18.6.6.2
Example of (*a*) a CNS module and (*b*) the corresponding module invocation. Input and output parameters are in bold. The module invocation is performed by specifying the '@' character, followed by the name of the module file and the module parameter substitutions. The ampersand (&) indicates that the particular symbol (*e.g.* '&fp') is substituted with the specified value in the invocation statement [*e.g.* 'fobs' in the case of '&fp' in (*b*)]. The module parameter substitution is performed literally, and any string of characters between the equal sign and the semicolon will be substituted.

The following example shows how the unit-cell parameters defined above (Fig. 18.6.3.1*b*) are passed into a module named 'compute_unit_cell_volume' (Fig. 18.6.6.1), which computes the volume of the unit cell from the crystal lattice parameters using well established formulae (Stout & Jensen, 1989):

$$\begin{aligned}&\texttt{@compute\_unit\_cell\_volume (cell} = \texttt{\&crystal\_lattice.unit\_cell;}\\ &\qquad\qquad\qquad\qquad \texttt{volume} = \texttt{\$cell\_volume;)}\end{aligned}\tag{18.6.6.1}$$

The parameter 'volume' is equated to the symbol '$cell_volume' upon invocation in order to return the result (the unit-cell volume) from this module. Note that the use of compound parameters to define the crystal lattice parameters (Fig. 18.6.3.1*b*) provides a convenient way to pass all required information into the module by referring to the base name of the compound

- Experimental phasing
  - Heavy-atom (Patterson) searches
  - Patterson refinement
  - Multiple isomorphous replacement phasing and site refinement
  - Multi-wavelength anomalous dispersion phasing and site refinement
- Molecular replacement
  - Patterson real-space and direct rotation searches
  - Patterson-correlation refinement
  - Fast FFT translation search
- Density modification
  - Creation of envelopes
  - Solvent flattening
  - Density averaging
  - Histogram matching
- Refinement
  - Maximum-likelihood targets
  - Torsion-angle molecular dynamics
  - Cartesian molecular dynamics
  - Conjugate-gradient minimization
  - Composite annealed omit map
- Other
  - Protein Data Bank deposition file generation
  - mmCIF file creation

Figure 18.6.7.1
Procedures and features available in CNS for structure determination by X-ray crystallography.

parameter (‘&crystal_lattice.unit_cell’) instead of having to specify each individual data element.

Fig. 18.6.6.2(*a*) shows another example of a CNS module: the module named ‘phase_distribution’ computes phase probability distributions using the Hendrickson & Lattman formalism (Hendrickson & Lattman, 1970; Hendrickson, 1979; Blundell & Johnson, 1976). An example for invoking the module is shown in Fig. 18.6.6.2(*b*). This module could be called from task files that need access to isomorphous phase probability distributions. It would be straightforward to change the module in order to compute different expressions for the phase probability distributions.

A large number of additional modules are available for crystallographic phasing and refinement. CNS library modules include space-group information, Gaussian atomic form factors, anomalous-scattering components, and molecular parameter and topology databases.

18.6.7. Task files

Task files consist of CNS language statements and module invocations. The CNS language permits the design and execution of nearly any numerical task in X-ray crystallographic structure determination using a minimal set of ‘hard-wired’ functions and routines. A list of the currently available crystallographic procedures and features is shown in Fig. 18.6.7.1.

Each task file is divided into two main sections: the initial parameter definition and the main body of the task file. The definition section contains definitions of all CNS parameters that are used in the main body of the task file. Modification of the main body of the file is not required, but may be done by experienced users in order to experiment with new algorithms.

```
{+ file: anneal.inp +}
{+ description: Crystallographic simulated annealing refinement +}
{+ authors: Axel T. Brunger, Luke M. Rice and Paul D. Adams +}
{+ reference: A.T. Brunger, J. Kuriyan and M. Karplus, Crystallographic
              R factor Refinement by Molecular Dynamics, Science
              235, 458-460 (1987) +}
{+ reference: A.T. Brunger, A. Krukowski and J. Erickson, Slow-Cooling
              Protocols for Crystallographic Refinement by Simulated
              Annealing, Acta Cryst. A46, 585-593 (1990) +}
{- begin block parameter definition -} define(
{====================== crystallographic data ========================}
{* space group *}
{* use International Table conventions with subscripts substituted by
   parenthesis *}
{===>} sg="P2(1)2(1)2(1)";

{* unit cell *}
{===>} a=61.76;   {===>} b=40.73;  {===>} c=26.74;
{===>} alpha=90;  {===>} beta=90;  {===>} gamma=90;

{* anomalous f' f'' library file *}
{* should be used when refining against anomalous data -
   libraries: "CNS_XTALLIB:anom_cu.lib" and "CNS_XTALLIB:anom_mo.lib" or
              a user created file.
   If blank no anomalous contribution will be included in the refinement *}
{===>} anom_library="";

{* reflection file *}
{===>} ref="example.hkl";

{* reciprocal space array containing observed amplitudes: required *}
{===>} obs_f="f_native";

{* reciprocal space array containing sigma values for amplitudes: required *}
{===>} obs_sigf="s_native";

{* reciprocal space array containing test set for cross-validation: required *}
{===>} test_set="test";
{* refinement target *}
{* mlf: maximum likelihood target using amplitudes
   mli: maximum likelihood target using intensities
   mlhl: maximum likelihood target using amplitudes and phase probability
         distribution
   residual: standard crystallographic residual
   vector: vector residual
   mixed: (1-fom)*residual + fom*vector
   e2e2: correlation coefficient using normalized E^2
   e1e1: correlation coefficient using normalized E
   f2f2: correlation coefficient using F^2
   f1f1: correlation coefficient using F *}
{+ choice: "mlf" "mli" "mlhl" "residual" "vector" "mixed"
          "e2e2" "e1e1" "f2f2" "f1f1" +}
{===>} reftarget="mlf";
}  {- end block parameter definition -}
```

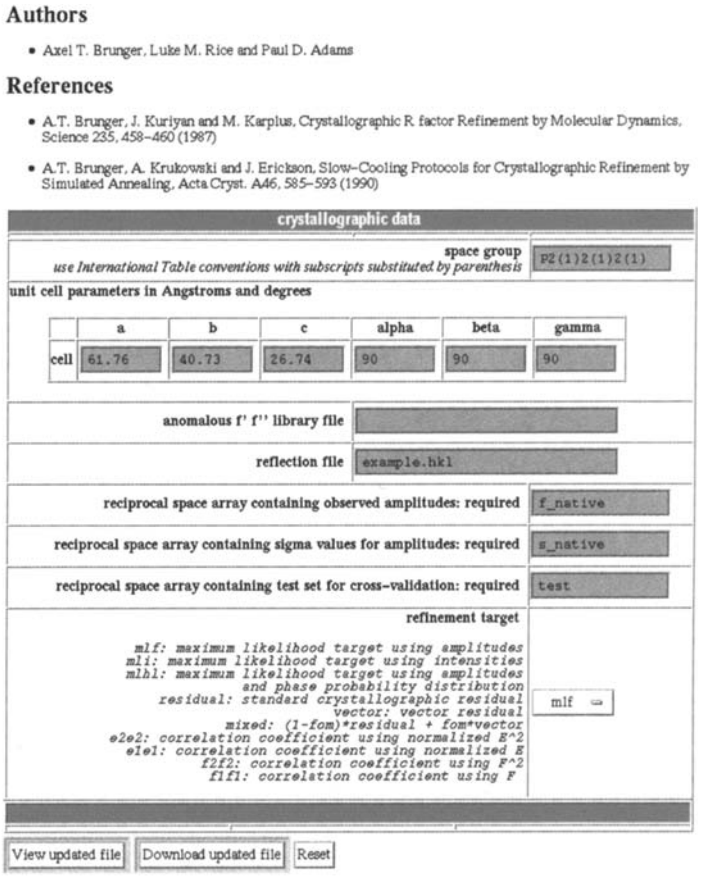
Figure 18.6.7.2
Example of a typical CNS task file: a section of the top portion of the simulated-annealing refinement protocol which contains the definition of various parameters that are needed in the main body of the task file. Each parameter is indicated by a name, an equal sign and an arbitrary sequence of characters terminated by a semicolon (*e.g.* ‘a=61.76;’). The top portion of each task file also contains commands for the HTML interface embedded in comment fields (indicated by braces, {...}). The commands that can be modified by the user in the HTML form are in bold.

The definition section also contains the directives that specify specific HTML features, *e.g.* text comments (indicated by {*...*}), user-modifiable fields (indicated by {=== >}), and choice boxes (indicated by {+ choice: ...+}). Fig. 18.6.7.2 shows a portion of the ‘define’ section of a typical CNS refinement task file.

The task files produce a number of output files (*e.g.* coordinate, reflection, graphing and analysis files). Comprehensive information about input parameters and results of the task are provided in these output files. In this way, the majority of the information required to reproduce the structure determination is kept with the results. Analysis data are often given in simple columns and rows of numbers. These data files can be used for graphing, for example, by using commonly available spreadsheet programs. An HTML graphical output feature for CNS which makes use of these analysis files is planned. In addition, list files are often produced that contain a synopsis of the calculation.

18.6.8. HTML interface

The HTML graphical interface uses HTML to create a high-level menu-driven environment for CNS (Fig. 18.6.8.1). Compact and



Authors

- Axel T. Brunger, Luke M. Rice and Paul D. Adams

References

- A.T. Brunger, J. Kuriyan and M. Karplus, Crystallographic R factor Refinement by Molecular Dynamics, Science 235, 458–460 (1987)
- A.T. Brunger, A. Krukowski and J. Erickson, Slow-Cooling Protocols for Crystallographic Refinement by Simulated Annealing, Acta Cryst. A46, 585–593 (1990)

crystallographic data

space group: P2(1)2(1)2(1)
use International Table conventions with subscripts substituted by parenthesis

unit cell parameters in Angstroms and degrees

	a	b	c	alpha	beta	gamma
cell	61.76	40.73	26.74	90	90	90

anomalous f' f'' library file:

reflection file: example.hkl

reciprocal space array containing observed amplitudes: required: f_native

reciprocal space array containing sigma values for amplitudes: required: s_native

reciprocal space array containing test set for cross-validation: required: test

refinement target: mlf

```
mlf: maximum likelihood target using amplitudes
mli: maximum likelihood target using intensities
mlhl: maximum likelihood target using amplitudes
          and phase probability distribution
residual: standard crystallographic residual
                     vector: vector residual
mixed: (1-fom)*residual + fom*vector
e2e2: correlation coefficient using normalized E^2
e1e1: correlation coefficient using normalized E
      f2f2: correlation coefficient using F^2
      f1f1: correlation coefficient using F
```

Figure 18.6.8.1
Example of a CNS HTML form page. This particular example corresponds to the task file in Fig. 18.6.7.2.

relatively simple Common Gateway Interface (CGI) conversion scripts are available that transform a task file into a form page and the edited form page back into a task file (Fig. 18.6.8.2). These conversion scripts are written in PERL.

A comprehensive collection of task files are available for crystallographic phasing and refinement (Fig. 18.6.7.1). New task files can be created or existing ones modified in order to address problems that are not currently met by the distributed collection of task files. The HTML graphical interface thus provides a common interface for distributed and 'personal' CNS task files (Fig. 18.6.8.2).

18.6.9. Example: combined maximum-likelihood and simulated-annealing refinement

CNS has a comprehensive task file for simulated-annealing refinement of crystal structures using Cartesian (Brünger *et al.*, 1987; Brünger, 1988) or torsion-angle molecular dynamics (Rice & Brünger, 1994). This task file automatically computes cross-validated σ_A estimates, determines the weighting scheme between the X-ray refinement target function and the geometric energy function (Brünger *et al.*, 1989), refines a flat bulk solvent model (Jiang & Brünger, 1994) and an overall anisotropic B value for the model by least-squares minimization, and subsequently refines the atomic positions by simulated annealing. Options are available for specification of alternate conformations, multiple conformers (Burling & Brünger, 1994), noncrystallographic symmetry constraints and restraints (Weis *et al.*, 1990), and 'flat' solvent models (Jiang & Brünger, 1994). Available target functions include the maximum-likelihood functions MLF, MLI and MLHL (Pannu & Read, 1996; Adams *et al.*, 1997; Pannu *et al.*, 1998). The user can choose between slow cooling (Brünger *et al.*, 1990) and constant-temperature simulated annealing, and the respective rate of cooling and length of the annealing scheme. For a review of simulated annealing in X-ray crystallography, see Brünger *et al.* (1997).

During simulated-annealing refinement, the model can be significantly improved. Therefore, it becomes important to recalculate the cross-validated σ_A error estimates (Kleywegt & Brunger, 1996; Read, 1997) and the weight between the X-ray diffraction target function and the geometric energy function in the course of the refinement (Adams *et al.*, 1997). This is important for the maximum-likelihood target functions that depend on the cross-validated σ_A error estimates. In the simulated-annealing task file, the recalculation of σ_A values and subsequently the weight for the crystallographic energy term are carried out after initial energy minimization, and also after molecular-dynamics simulated annealing.

18.6.10. Conclusions

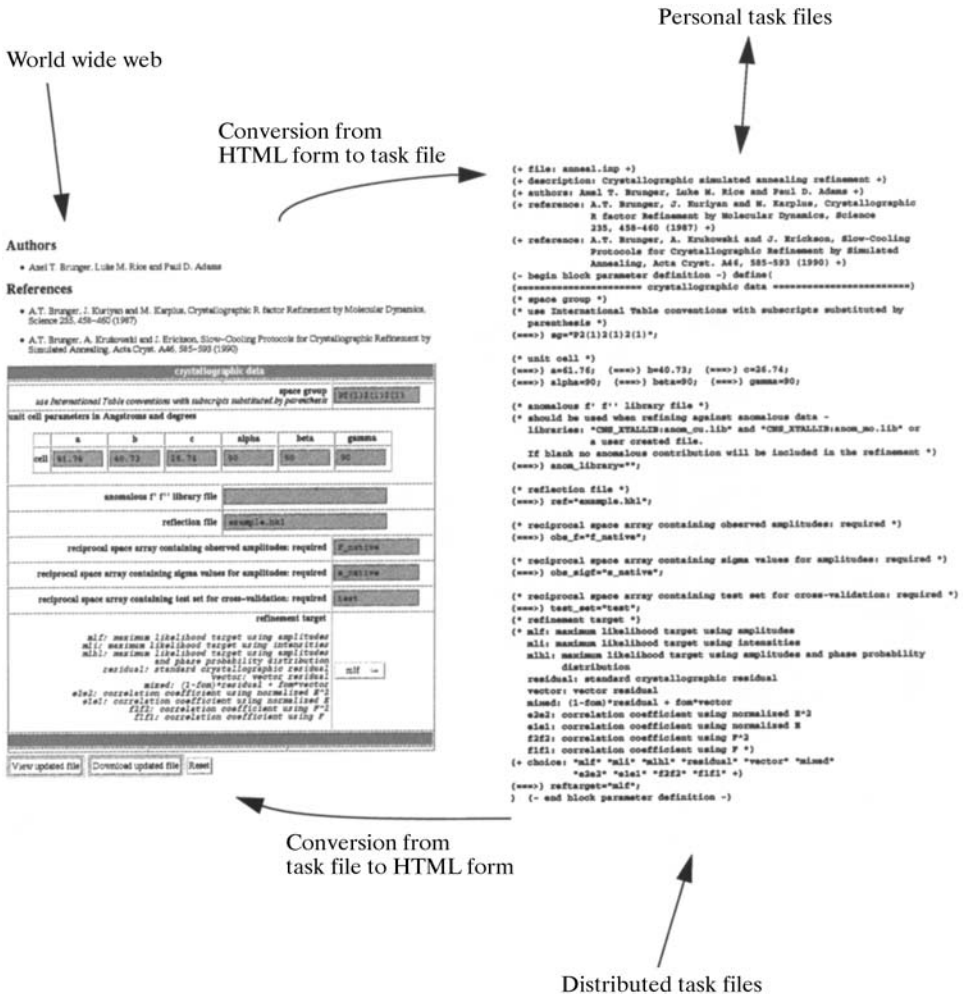
CNS is a general system for structure determination by X-ray crystallography and solution NMR. It covers the whole spectrum of methods used to solve X-ray or solution NMR structures. The multi-layer architecture allows use of the system with different levels of expertise. The HTML interface allows the novice to perform standard tasks. The interface provides a convenient means of editing complicated task files, even for the expert (Fig. 18.6.8.2). This graphical interface makes it less likely that an important parameter will be overlooked when editing the file. In addition, the graphical interface can be used with any task file, not just the standard distributed ones. HTML-based documentation and graphical output are planned in the future.

Most operations within a crystallographic algorithm are defined through modules and task files. This allows for the development of new algorithms and for existing algorithms to be precisely defined and easily modified without the need for source-code modifications.

The hierarchical structure of CNS allows extensive testing at each level. For example, once the source code and CNS basic commands have been tested, testing of the modules and task files is performed. A test suite consisting of more than a hundred test cases is frequently evaluated during CNS development in order to detect and correct programming errors. Furthermore, this suite is run on several hardware platforms in order to detect any machine-specific errors. This testing scheme makes CNS highly reliable.

Algorithms can be readily understood by inspecting the modules or task files. This self-documenting feature of the modules provides a powerful teaching tool. Users can easily interpret an algorithm and compare it with published methods in the literature. To our knowledge, CNS is the only system that enables one to define symbolically any target function for a broad range of applications, from heavy-atom phasing or molecular-replacement searches to atomic resolution refinement.

Support by the Howard Hughes Medical Institute and the National Science Foundation to ATB (DBI-9514819 and ASC 93-181159), the Natural Sciences and Engineering Research Council of Canada to NSP, the Howard Hughes Medical Institute and the Medical Research Council of Canada to RJR (MT11000), the Netherlands Foundation for Chemical Research (SON–NWO) to



Authors

- Axel T. Brunger, Luke M. Rice and Paul D. Adams

References

- A.T. Brunger, J. Kuriyan and M. Karplus, Crystallographic R factor Refinement by Molecular Dynamics, Science 235, 458–460 (1987)
- A.T. Brunger, A. Krukowski and J. Erickson, Slow-Cooling Protocols for Crystallographic Refinement by Simulated Annealing, Acta Cryst. A46, 585–593 (1990)

crystallographic data

space group: P2(1)2(1)2(1)
use International Table conventions with subscripts substituted by parenthesis

unit cell parameters in Angstroms and degrees

	a	b	c	alpha	beta	gamma
cell	61.76	40.73	26.74	90	90	90

anomalous f' f'' library file:
reflection file: example.hkl
reciprocal space array containing observed amplitudes: required: f_native
reciprocal space array containing sigma values for amplitudes: required: s_native
reciprocal space array containing test set for cross-validation: required: test
refinement target: mlf

```
mlf: maximum likelihood target using amplitudes
mli: maximum likelihood target using intensities
mlhl: maximum likelihood target using amplitudes
and phase probability distribution
residual: standard crystallographic residual
vector: vector residual
mixed: (1-fom)*residual + fom*vector
e2e2: correlation coefficient using normalized E^2
e1e1: correlation coefficient using normalized E
f2f2: correlation coefficient using F^2
f1f1: correlation coefficient using F
```

```
{+ file: anneal.inp +}
{+ description: Crystallographic simulated annealing refinement +}
{+ authors: Axel T. Brunger, Luke M. Rice and Paul D. Adams +}
{+ reference: A.T. Brunger, J. Kuriyan and M. Karplus, Crystallographic
              R factor Refinement by Molecular Dynamics, Science
              235, 458-460 (1987) +}
{+ reference: A.T. Brunger, A. Krukowski and J. Erickson, Slow-Cooling
              Protocols for Crystallographic Refinement by Simulated
              Annealing, Acta Cryst. A46, 585-593 (1990) +}
{- begin block parameter definition -} define(
{====================== crystallographic data ========================}
{* space group *}
{* use International Table conventions with subscripts substituted by
   parenthesis *}
{===>} sg="P2(1)2(1)2(1)";

{* unit cell *}
{===>} a=61.76;  {===>} b=40.73;  {===>} c=26.74;
{===>} alpha=90;  {===>} beta=90;  {===>} gamma=90;

{* anomalous f' f'' library file *}
{* should be used when refining against anomalous data -
   libraries: "CNS_XTALLIB:anom_cu.lib" and "CNS_XTALLIB:anom_mo.lib" or
              a user created file.
   If blank no anomalous contribution will be included in the refinement *}
{===>} anom_library="";

{* reflection file *}
{===>} ref="example.hkl";

{* reciprocal space array containing observed amplitudes: required *}
{===>} obs_f="f_native";

{* reciprocal space array containing sigma values for amplitudes: required *}
{===>} obs_sigf="s_native";

{* reciprocal space array containing test set for cross-validation: required *}
{===>} test_set="test";
{* refinement target *}
{* mlf: maximum likelihood target using amplitudes
   mli: maximum likelihood target using intensities
   mlhl: maximum likelihood target using amplitudes and phase probability
         distribution
   residual: standard crystallographic residual
   vector: vector residual
   mixed: (1-fom)*residual + fom*vector
   e2e2: correlation coefficient using normalized E^2
   e1e1: correlation coefficient using normalized E
   f2f2: correlation coefficient using F^2
   f1f1: correlation coefficient using F *}
{+ choice: "mlf" "mli" "mlhl" "residual" "vector" "mixed"
           "e2e2" "e1e1" "f2f2" "f1f1" +}
{===>} reftarget="mlf";
)  {- end block parameter definition -}
```

Figure 18.6.8.2
Use of the CNS HTML form page interface, emphasizing the correspondence between input fields in the form page and parameters in the task file.

PG and the Howard Hughes Medical Institute to LMR is gratefully acknowledged.

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, pp. 375–377. London: Academic Press.

Bricogne, G. (1984). *Maximum entropy and the foundations of direct methods. Acta Cryst.* A**40**, 410–445.

Brünger, A. T. (1988). *Crystallographic refinement by simulated annealing: application to a 2.8 Å resolution structure of aspartate aminotransferase. J. Mol. Biol.* **203**, 803–816.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography & NMR System (CNS): a new software suite for macromolecular structure determination. Acta Cryst.* D**54**, 905–921.

Brünger, A. T., Adams, P. D. & Rice, L. M. (1997). *New applications of simulated annealing in X-ray crystallography and solution NMR. Structure*, **5**, 325–336.

Brünger, A. T., Karplus, M. & Petsko, G. A. (1989). *Crystallographic refinement by simulated annealing: application to crambin. Acta Cryst.* A**45**, 50–61.

Brünger, A. T., Krukowski, A. & Erickson, J. W. (1990). *Slow-cooling protocols for crystallographic refinement by simulated annealing. Acta Cryst.* A**46**, 585–593.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Burling, F. T. & Brünger, A. T. (1994). *Thermal motion and conformational disorder in protein crystal structures: comparison of multi-conformer and time-averaging models. Isr. J. Chem.* **34**, 165–175.

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Direct observation of protein solvation and discrete disorder with experimental crystallographic phases. Science*, **271**, 72–77.

Graham, I. S. (1995). *The HTML Sourcebook.* John Wiley and Sons.

Hendrickson, W. A. (1979). *Phase information from anomalous-scattering measurements. Acta Cryst.* A**35**, 245–247.

Hendrickson, W. A. (1991). *Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science*, **254**, 51–58.

Hendrickson, W. A. & Lattman, E. E. (1970). *Representation of phase probability distributions for simplified combination of independent phase information. Acta Cryst.* B**26**, 136–143.

Jiang, J.-S. & Brünger, A. T. (1994). *Protein hydration observed by X-ray diffraction: solvation properties of penicillopepsin and neuraminidase crystal structures. J. Mol. Biol.* **243**, 100–115.

Kleywegt, G. J. & Brünger, A. T. (1996). *Checking your imagination: applications of the free R value. Structure*, **4**, 897–904.

Otwinowski, Z. (1991). In *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.

Read, R. J. (1994). *Maximum likelihood refinement of heavy atoms.* Lecture notes for a workshop on isomorphous replacement methods in macromolecular crystallography. American Crystallographic Association Annual Meeting, 1994, Atlanta, GA, USA.

Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol.* **277**, 110–128.

Rice, L. M. & Brünger, A. T. (1994). *Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. Proteins Struct. Funct. Genet.* **19**, 277–290.

Stout, G. H. & Jensen, L. H. (1989). *X-ray Structure Determination*, p. 33. New York: Wiley Interscience.

Weis, W. I., Brünger, A. T., Skehel, J. J. & Wiley, D. C. (1990). *Refinement of the influenza virus haemagglutinin by simulated annealing. J. Mol. Biol.* **212**, 737–761.

International Tables for Crystallography (2012). Vol. F, Chapter 18.7, pp. 520–524.

Chapter 18.7. The *TNT* refinement package

D. E. Tronrud and L. F. Ten Eyck

18.7.1. Scope and function of the package

TNT (Tronrud *et al.*, 1987) is a computer program package that optimizes the parameters of a molecular model given a set of observations and indicates the location of errors that it cannot correct. Its authors presume the principal set of observations to be the structure factors observed in a single-crystal diffraction experiment. To complement such a data set, which for most macromolecules has limitations, stereochemical restraints such as standard bond lengths and angles are also used as observations.

A molecule is parameterized as a set of atoms, each with a position in space, an isotropic B factor and an occupancy. The complete model also includes an overall scale factor, which converts the arbitrary units of the measured structure factors to e Å^{-3}, and a two-parameter model of the electron density of the bulk solvent.

Because a *TNT* model of a macromolecule does not allow anisotropic B factors, *TNT* cannot be used to finish the refinement of any structure that diffracts to high enough resolution to justify the use of these parameters. If one has a crystal that diffracts to 1.4 Å or better, the final model should probably include these parameters and *TNT* cannot be used. One may still use *TNT* in the early stages of such a refinement because one usually begins with only isotropic B's.

At the other extreme of resolution, *TNT* begins to break down with data sets limited to only about 3.5 Å data. This breakdown occurs for two reasons. First, at 3.5 Å resolution, the maps can no longer resolve β-sheet strands or α-helices. The refinement of a model against data of such low resolution requires strong restraints on dihedral angles and hydrogen bonds – tasks for which *TNT* is not well suited. Second, the errors in an initial model constructed with only 3.5 Å data are usually of such a magnitude and quality that the function minimizer in *TNT* cannot correct them.

18.7.2. Historical context

The design of *TNT* began in the late 1970s, and the first publishable models were generated by *TNT* in 1981 (Holmes & Matthews, 1981). Its design was greatly influenced by observations of the strength and weaknesses of programs then available.

The first refinement of a protein model was performed by Jensen and co-workers at the University of Washington (Watenpaugh *et al.*, 1973). This structure refinement was atypical because of the availability of high-resolution data. The techniques of pre-least-squares small-molecule refinement were simply applied to this much larger model. Since many of the calculations were performed manually, no comprehensive software package was created for distribution.

It was quickly realized that for macromolecular refinement to become common, the calculations had to be fully automated and ideal stereochemistry had to be enforced. In the late 1970s, four programs became available, all of which automated the refinement calculations, but each implemented the enforcement of stereochemistry in different ways. They were *PROLSQ* (Hendrickson & Konnert, 1980), *EREF* (Jack & Levitt, 1978), *CORELS* (Sussman *et al.*, 1977) and *FFTSF* (Agarwal, 1978). *PROLSQ* was, ultimately, the most popular.

At one end of the spectrum lay *FFTSF*. This program optimized its models to the diffraction data while completely ignoring ideal geometry. Following a number of iterations of optimizing the fit of the model to the structure factors, the geometry was idealized by running a separate program. At the other extreme was *CORELS*. It optimized its models to the diffraction data while allowing no deviations from ideal stereochemistry. The model was allowed to change only through the rotation of single bonds and the movement of rigid groups. Both approaches were frustrating to a certain extent. With *FFTSF* it was a struggle to find a model that agreed with all observations at once. With *CORELS* it was difficult to get the model to fit the density, because small and, apparently, insignificant deviations from ideality often added up after many residues to large and significant displacements, and these were forbidden. Neither approach to stereochemistry seemed very convenient, although *CORELS* was used for early-stage refinement for many years because of its exceptional radius of convergence.

Both *PROLSQ* and *EREF* enforced ideal stereochemistry and agreement with the diffraction data simultaneously. This strategy proved very convenient and generated models that satisfied their users. The two programs differed significantly in the form in which they required the ideal values be entered. *PROLSQ* required that the ideal values for both bond lengths and bond angles be entered as distances, *e.g.* an angle was defined by the distance between the two extreme atoms. *EREF* required that the standard value for an angle simply be entered as the number of degrees. Since *EREF* stored its library of standard values in the same terms as those with which people were familiar, it was much easier to enter the values.

These two programs differed in another way as well. *PROLSQ* stored ideal values for the stereochemistry of each type of residue (*e.g.* alanine, glycine *etc.*), while *EREF* parameterized the library in terms of atom types. For example, the angle formed by three atoms, the first a keto oxygen, the second a carbonyl carbon and the third an amide nitrogen, would have a particular ideal value regardless of where these three atoms occurred. In this matter, *PROLSQ* was more similar to the thought patterns of crystallographers.

18.7.3. Design principles

TNT was designed with three fundamental principles in mind. Each principle has a number of consequences that shaped the ultimate form of the package.

18.7.3.1. Refinement should be simple to run

The user should not be burdened with the choice of input parameters that they may not be qualified to choose. They also should not be forced to construct an input file that is obscure and difficult to understand. It is hard now to remember what most

Copyright © 2012 International Union of Crystallography

computer programs were like in the 1970s. Usually, the input to the program was a block of numbers and flags where the meaning of each item was defined by its line and column numbers. This block not only contained information the programmer could never anticipate, like the cell constants, but defined how the computer's memory should be allocated and obscure parameters that could only be estimated after careful reading of research papers.

TNT was one of the first programs in crystallography to have its input introduced with keywords and to allow input statements to come in any order. As an example of the difference, consider the resolution limits. Usually, a crystallographic program would have a line in its input similar to

```
99.0, 1.9,
```

One had to recognize this line amongst many as the line containing the resolution limits. (In many programs, a value of 99 was used to indicate that no lower-resolution limit was to be applied.) In *TNT* the same data would be entered as

```
RESOLUTION 1.9
```

The keyword identifies the data as the resolution limit(s). If the statement contained two numbers, they were considered the upper and lower limits of the diffraction data.

The preceding example also shows how default values can be implemented by a program much more safely with keyword-based input. In the previous scheme, if a value was ever to be changed by the user, its place had to be allocated in the input block. This often left numbers floating in the block which were almost never changed, and because they were so infrequently referred to, they were usually unrecognized by the user. It was quite possible for one of these numbers to be accidentally changed and the error unnoticed for quite some time. When the data are introduced with keywords, a data item is not mentioned if the default value is suitable.

18.7.3.2. Refinement should run quickly and use as little memory as possible

The most time-consuming calculations in refinement are the calculation of structure factors from atomic coordinates and the calculation of derivatives of the part of the residual dependent upon the diffraction data with respect to the atomic parameters. The quickest means of performing these calculations requires the use of space-group-optimized fast Fourier transforms (FFTs). The initial implementation of *TNT* used FFTs to calculate structure factors, but the much slower direct summation method to calculate the derivatives. Within a few years, Agarwal's method (Agarwal, 1978; Agarwal *et al.*, 1981) was incorporated into *TNT* and from then on all crystallographic calculations were performed with FFTs.

The FFT programs of Ten Eyck (1973, 1977) made very efficient use of computer memory. Another means of saving memory was to recognize that the code for calculating stereochemical restraints did not need to be in the memory when the crystallographic calculations were being performed and *vice versa*. There were two ways to save memory using this information. One could create a series of 'overlays' or one could break the calculation into a series of separate programs. The means for defining an overlay structure were never standardized and could not be ported from one type of computer to another and were, therefore, never attempted in *TNT*. For this reason, and a number of others mentioned here, *TNT* is not a single program but a collection of programs, each with a well defined and specialized purpose.

18.7.3.3. The source code should not require customization for each project

The need to state this goal seems remarkable in these modern times, but the truth is that most computer programs in the 1970s required specific customizations before they could be used. The simplest modifications were the definitions of the maximum number of atoms, residues, atom types *etc.* accepted by the program. These modifications are still required in Fortran77 programs because that language does not allow the dynamic allocation of memory. However, in most programs today the limits are set high enough that the standard configuration does not present a problem.

The most difficult modification required for programs like *PROLSQ* was to adapt the calculations to the space group in hand. Their authors usually included code for the space groups they were particularly interested in, leaving all others to be implemented by the user. Writing code for a new space group was often a daunting task for someone who was not an expert programmer and had no tools for testing the modifications.

It is too burdensome to require the user to understand sufficiently the internal workings of a complex calculation that they can code and debug central subroutines of a refinement program. In its initial implementation, *TNT* avoided this problem, to an extent, by performing the space-group-specific calculations in separate programs. At least the user did not need to modify an existing program. All that was required was the construction of a program that read the proper format file, performed the calculation and wrote its answer in the proper format. The user was required to supply both a program that could calculate structure factors from the model and another program that could calculate the derivative of the diffraction component of the residual function with respect to the atomic parameters of the model.

While a structure-factor program could usually be located, either by finding an existing program or by expanding the model to a lower-symmetry space group for which a program did exist, the requirement of creating a derivative program proved too great a burden. The derivation of the space-group-specific calculation, its implementation and debugging proved too difficult for almost everyone, and this design was quickly abandoned. Instead, an implementation of Agarwal's (1978) algorithm was created. In this method, the derivatives are calculated with a series of convolutions with an $F_o - F_c$ map. The calculation of the map is the only space-group-specific part of the calculation, and this was done with a separate program for calculating Fourier syntheses. Such programs were as easy to come by as structure-factor calculation programs and could be replaced by a lower-symmetry program if required.

While it is easier to find or write a program that only calculates a Fourier transform and much easier to debug one than to debug a modification to a larger and more complex program, it is still difficult. The lack of availability of programs for the space group of a crystal often prevented the use of *TNT*. Over time, programs for more space groups were written and distributed with *TNT*. Eventually, a method was developed by one of *TNT*'s authors in which FFTs could be calculated using a single program as efficiently as the original space-group-specific programs. Once this program existed, there was no longer the need for isolated structure-factor and Fourier synthesis programs. These calculations have disappeared into the heart of *TNT*, and *TNT* consists of many fewer programs today than in the past.

18.7.4. Current structure of the package

TNT presents different faces to different users. Some users simply want to run refinement; they see the shell interface. Others want to use the *TNT* programs in untraditional ways; they see the program interface. A few users want to change the basic calculations of *TNT*; they see the library interface.

The shell interface is the view of *TNT* that most people see. It is the most recent structural addition, having been added in release 5E in 1995. At this level, the restraints, weights and parameters of the model are described in the '*TNT* control file' and the user performs particular calculations by giving commands at the shell prompt. For example, refinement is performed with the 'tnt' command and maps are calculated for examination with some graphics program with the 'make_maps' command. *TNT* is supplied with about two dozen shell commands. These commands allow the running of refinement, the conversion of the model to and from *TNT*'s internal format, and the examination of the model to locate potential problem spots. The *TNT Users' Guide* describes the use of *TNT* at this level.

The program interface consists of the individual *TNT* programs along with their individual capabilities. *TNT* consists of the program *Shift*, which handles all the minimization calculations, a program for each module (restraints that fall into a common class, *e.g.* diffraction data, ideal stereochemistry and noncrystallographic symmetry) and a number of utility programs of which the most important member is the program *Convert*, which reads and writes coordinate files in many formats. The user can write shell scripts (or modify those supplied with *TNT*) to perform a great many tasks that cannot be accessed with the standard set of scripts. The *TNT Reference Manual* describes the operation of each program.

If the programs in *TNT* do not perform the calculation wanted, the source code can be modified. The source code to *TNT* is supplied with the standard distribution. In order to make the code more manageable and understandable, it is divided into half a dozen libraries. All *TNT* programs use the lowest-level library to ensure consistency of the 'look and feel' and use the basic data structures for storage of the model's parameters and the vital crystal data. To add new functionality, one can either modify an existing program, write a new program using the *TNT* libraries as a start, or write a new program from scratch ignoring the *TNT* libraries. As long as a program can read and write files of the same format as the rest of *TNT*, it will work well with *TNT*, even if it does not share any code. A library exists, but is not copyrighted, that contains subroutines to read and write the crystallographic file formats used by the rest of *TNT*.

18.7.5. Innovations first introduced in *TNT*

TNT was not only designed to be an easy-to-use tool for the refinement of macromolecular models, but also as a tool for testing new ideas in refinement. Since its source code is designed to allow easy reordering of tasks and simple modifications, a number of innovations in refinement made their first appearance in *TNT*. These features include the following.

18.7.5.1. Identifying and restraining symmetry-related contacts (1982)

Without a search for symmetry-related bad contacts, it was quite common to build atoms into the same density from two different sides of the molecule. A number of models in the Protein Data Bank contain these types of errors because neither the refinement nor the graphics programs available at that time would indicate this type of error.

18.7.5.2. The ability of a single package to perform both individual atom and rigid-body refinement (1982)

Prior to *TNT*, one often started a refinement with rigid-body refinement using *CORELS* and then switched to another program. *TNT* was the first refinement package to allow both styles of refinement. One was not required to learn about two different packages when running *TNT*.

18.7.5.3. Space-group optimized FFTs for all space groups (1989)

This innovation allowed *TNT* to run efficiently in all space groups available to macromolecular crystals.

18.7.5.4. Modelling bulk solvent scattering via local scaling (~1989)

With a simple and quick model of the scattering of the bulk solvent in the crystal (Tronrud, 1997), the low-resolution data could be used in refinement for the first time. The inclusion of these data in the calculation of maps greatly improved their appearance.

18.7.5.5. Preconditioned conjugate-gradient minimization (1990)

This method of minimization (Axelsson & Barker, 1984; Tronrud, 1992) allows the direct inclusion of the diagonal elements of the second-derivative matrix and the indirect inclusion of its off-diagonal elements. An additional benefit is that it allows both positional parameters and B factors to be optimized in each cycle. Previously, one was required to hold one class of parameter fixed while the other was optimized. It is much more efficient and simpler for the user to optimize all parameters at once. This method, because it incorporates the diagonal elements directly, produces sets of B factors that agree with the diffraction data better than those from the simple conjugate-gradient method.

18.7.5.6. Restraining stereochemistry of chemical links to symmetry-related molecules (~1992)

It is not uncommon for crystallization enhancers to be found on a special position in the crystal. In addition, cross-linking the molecules in a crystal is often done for various reasons. In both cases, the model contains chemical bonds to a molecule or atoms in another asymmetric unit of the crystal. In order for the stereochemistry of these links to be properly restrained, it must be possible to describe such a link to the refinement program.

18.7.5.7. Knowledge-based B-factor restraints (~1994)

When the resolution of the diffraction data set is less than about 2 Å, the individual B factors of a refined model are observed to vary wildly from atom to atom, even when the atoms are bonded to one another. This pattern is not reasonable if one interprets the B factor as a measure of the vibrational motion of the atom. Traditionally, one applies an additional restraint on the B factors of the model, where the ideal value for the difference in B factor for two bonded atoms is zero.

Since it is clear from examinations of higher-resolution models that the *B* factors generally increase from one side of a bond to the other (*e.g.* moving from the main chain to the end of a side chain), the traditional restraint is flawed. A restraint library was generated (Tronrud, 1996) where each bond in a residue is assigned a preferred increment in *B* factor and a confidence (standard deviation) in that increment.

18.7.5.8. Block-diagonal preconditioned conjugate-gradient minimization with pseudoinverses (1998)

With this enhancement, *TNT*'s minimizer treats the second-derivative matrix as a collection of 5×5 element blocks along its diagonal, one block for each atom. While this method improves the rate of convergence for noncrystallographic symmetry restraints, its most significant feature is that it allows the refinement of atoms located on special positions without special handling by the user.

18.7.5.9. Generalization of noncrystallographic symmetry operators to include shifts in the average B factor (1998)

It is rather common in crystals containing multiple copies of a molecule in the asymmetric unit for one or more molecules to have a higher *B* factor than the others. If the transformation that generates each copy of the molecule consists only of a rotation and translation of the positions of the atoms, the difference in *B* factors cannot be modelled. The transformations used in *TNT* now consist of a rotation, translation, a *B*-factor shift and an occupancy shift.

18.7.6. *TNT* as a research tool

TNT was intended not only as a tool for performing refinement, but as a tool for developing new ideas in refinement. While most of the latter has been done by *TNT*'s authors, several others have made good use of *TNT* in this fashion. If one has an idea to test, the overhead of writing an entire refinement package to perform that test is overwhelming. *TNT* allows modification at a number of levels, so one can choose to work at the level that allows the easiest implementation of the idea. Several examples follow.

18.7.6.1. Michael Chapman's real-space refinement package

At Florida State University, Chapman has implemented a real-space refinement package, principally intended for the refinement of virus models, using *TNT*. He was able to use *TNT*'s minimizer and stereochemical restraints unchanged along with programs he developed to implement his method. More information about this package can be found at http://www.sb.fsu.edu/~rsref.

18.7.6.2. Randy Read's maximum-likelihood function

When Navraj Pannu wanted to implement Read's maximum-likelihood refinement functions (Pannu & Read, 1996) in *TNT*, he chose not to implement it as a separate program, but modified *TNT*'s source code to create a new version of the program *Rfactor*, named *Maxfactor*.

18.7.6.3. J. P. Abrahams' likelihood-weighted noncrystallographic symmetry restraints

Abrahams (1996) conceived the idea that because some amino-acid side chains can be expected to violate the noncrystallographic symmetry (NCS) of the crystal more than others, one could develop a library of the relative strength with which each atom of each residue type would be held by the NCS restraint. He chose to determine these strengths from the average of the current agreement to the NCS of all residues of the same type. For example, if the lysine side chains do not agree well with their NCS mates, the NCS will be loosely enforced for those side chains. On the other hand, if almost all the valine side chains agree well with their mates, then the NCS will be strongly enforced for the few that do not agree well.

He chose to implement this idea by modifying the source code for the *TNT* program *NCS*. Since the calculations involved in implementing this idea are simple, the extent of the modification was not large.

18.7.6.4. The Buster refinement package

Bricogne & Irwin (1996) have developed a maximum-likelihood refinement package using *TNT*. Not only are *TNT*'s minimizer and stereochemical restraints used, but many of the calculations of the maximum-likelihood residual's derivatives are performed using *TNT* programs. While Bricogne and co-workers have not needed to modify *TNT* programs to implement their ideas, there is ongoing collaboration between them and *TNT*'s authors on the development of commands that allow access to some previously internal calculations. More information about *Buster* can be found at http://www.globalphasing.com/buster/.

18.7.7. Current status of *TNT*

TNT is no longer being actively developed at the University of Oregon. The principal use of *TNT* today is from within the *Buster* package of Global Phasing, Ltd. Over the last ten years, they have made considerable modifications to their branch of the source code which have greatly increased the radius and rate of convergence of the optimizer and expanded the scope of the distributed library of standard groups in the ideal stereochemistry library. These enhancements are only available *via* their distribution of *Buster*.

References

Abrahams, J. P. (1996). *Likelihood-weighted real space restraints for refinement at low resolution*. In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey. Warrington: Daresbury Laboratory.

Agarwal, R. C. (1978). *A new least-squares technique based on the fast Fourier transform algorithm. Acta Cryst.* A**34**, 791–809.

Agarwal, R. C., Lifchitz, A. & Dodson, E. (1981). *Block diagonal least squares refinement using fast Fourier techniques*. In *Refinement of Protein Structures*, edited by P. A. Machin, J. W. Campbell & M. Elder. Warrington: Daresbury Laboratory.

Axelsson, O. & Barker, V. (1984). *Finite element solution of boundary value problems*, ch. 1, pp. 1–63. Orlando: Academic Press.

Bricogne, G. & Irwin, J. J. (1996). *Maximum-likelihood refinement of incomplete models with BUSTER + TNT*. In *Proceedings of the Macromolecular Crystallographic Computing School*, edited by P. Bourne & K. Watenpaugh. http://www.iucr.org/resources/commissions/crystallographic-computing/schools/school96.

Hendrickson, W. A. & Konnert, J. H. (1980). *Incorporation of stereochemical information into crystallographic refinement*. In *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.23. Bangalore: Indian Academy of Sciences.

Holmes, M. A. & Matthews, B. W. (1981). *Binding of hydroxamic acid inhibitors to crystalline thermolysin suggests a pentacoordinate zinc intermediate in catalysis. Biochemistry*, **20**, 6912–6920.

Jack, A. & Levitt, M. (1978). *Refinement of large structures by simultaneous minimization of energy and R factor. Acta Cryst.* A**34**, 931–935.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood*. In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey. Warrington: Daresbury Laboratory.

Sussman, J. L., Holbrook, S. R., Church, G. M. & Kim, S.-H. (1977). *A structure-factor least-squares refinement procedure for macromolecular structures using constrained and restrained parameters. Acta Cryst.* A**33**, 800–804.

Ten Eyck, L. F. (1973). *Crystallographic fast Fourier transforms. Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* A**33**, 486–492.

Tronrud, D. E. (1992). *Conjugate-direction minimization: an improved method for the refinement of macromolecules. Acta Cryst.* A**48**, 912–916.

Tronrud, D. E. (1996). *Knowledge-based B-factor restraints for the refinement of proteins. J. Appl Cryst.* **29**, 100–104.

Tronrud, D. E. (1997). *TNT refinement package*. *Methods Enzymol.* **277**, 306–319.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *An efficient general-purpose least-squares refinement program for macromolecular structures. Acta Cryst.* A**43**, 489–501.

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Refinement of the model of a protein: rubredoxin at 1.5 Å resolution. Acta Cryst.* B**29**, 943–956.

International Tables for Crystallography (2012). Vol. F, Chapter 18.8, pp. 525–528.

Chapter 18.8. *ARP/wARP* – automated model building and refinement

V. S. Lamzin, A. Perrakis and K. S. Wilson

A macromolecular model is an interpretation of the end result of an X-ray experiment – the *electron-density map* – and is aimed to be consistent with both the experimental data and stereochemical knowledge. At the start of the process, the crystallographer must construct an initial molecular model in the electron density. In the early days this was done literally with nuts and bolts, but was subsequently carried out with the aid of interactive modelling software. Over the last decade, this time-demanding and highly subjective step, heavily dependent on the researcher's experience, has been largely substituted by automated procedures. Automated model building programs, of which *ARP/wARP* was one of the first to be widely used, deliver high-quality molecular models rapidly and efficiently from a sufficiently good set of initial phases.

18.8.1. Refinement and model building are two parts of modelling a structure

The conventional view of crystallographic refinement of macromolecules is the optimization of the parameters of a model to fit both the experimental data and a set of *a priori* stereochemical observations. The user provides a model, whose parameters are allowed to vary during a set minimization cycles based on least-squares or maximum-likelihood protocols. In conventional refinement procedures, the presence of the atoms is fixed, *i.e.* there is no addition or removal of parts of the model. As a result, users are often faced with a situation where several atoms in the model lie in one place, while the density maps suggest an entirely different location. The minimization procedures themselves are unable to make such major modifications to the model. Manual intervention, consisting of moving atoms to a more appropriate place using molecular graphics, density maps and geometrical assumptions, is required to enable improvement of the model to proceed.

ARP/wARP has built on two innovative concepts (Lamzin & Wilson, 1993, 1997; Perrakis *et al.*, 1999).

(1) It challenged the previous separation of model building and refinement software by taking a more general view of the underlying phase optimization through coordinate manipulation and by incorporating the data-driven extension of the macromolecular model as part of an integrated approach. Addition and/or deletion of atoms (*model update*) and complete re-evaluation of the model with an aim of creating a modified one that better describes the electron density (*model reconstruction*) are essential components. *Atom removal* is primarily based on the interpolated likelihood-weighted electron density at an atomic centre and the agreement of the atomic density distribution with a target shape. *Atom addition* uses the difference likelihood-gradient synthesis. The selection is based on map grid points rather than peaks, as the latter are often poorly defined and may overlap with neighbouring peaks or existing atoms, especially if the resolution and phases are poor.

(2) *ARP/wARP* also challenged the view that throughout the structure-determination process a macromolecular model should always consist of a known set of atoms of an appropriate type, connected by bonds according to the known stereochemistry. The concept of a *hybrid model*, described in detail in Section 18.8.2, was introduced where a partial macromolecular model is deduced from the interpretable part of the electron density, while the remaining regions are modelled with free atoms that have undefined chemical identity and no bonds. This placed structure solution on a more unified foundation and enabled automation in linking the entire procedure.

The hybrid model undergoes optimization with the *REFMAC* engine (Murshudov *et al.*, 1997), where the model parameters are adjusted to fit better the experimental data and *a priori* stereochemical expectations. If the quality of the hybrid model is sufficiently high, the phases improve overall and a more complete model can be constructed from the enhanced electron density. The hybrid model iteratively evolves (precisely what is needed to escape from a local minimum of the optimization landscape) and should converge on the final macromolecular model, with the remaining free atoms approximating the surrounding solvent structure. The fundamental limitations of *ARP/wARP* as well as related macromolecular structure determination approaches are the amount of the available X-ray data (resolution) and the quality of the initial phase estimates.

18.8.2. Free-atom and hybrid models

A crystallographic electron-density map is always sampled on a regular grid. The goal of macromolecular model building is to condense the electron-density information to a molecular model. *ARP/wARP* entirely follows the paradigm of syntactic model construction starting from free atoms, through candidate Cα atoms, peptide and dipeptide units and chain fragments to sequence docking and side-chain building. As a first step, *ARP/wARP* condenses the map information to a set of *free atoms* (Agarwal & Isaacs, 1977), as described by Lamzin & Wilson (1997), that have undefined chemical identity: these atoms are chosen to represent the electron density as accurately as possible, but resemble a protein-like model in their distribution. The free atoms can be complemented by an initial (partial) protein model if available. To build a free-atom model, only an estimate of the molecular weight of the protein is required, without any sequence information. For example, for protein-model building a map covering a crystallographic asymmetric unit on a fine grid of about 0.33 Å is constructed. The model is gradually expanded from a seed by the stepwise addition of free atoms in lower electron density and at bonding distances from existing atoms. The procedure continues until the number of atoms is about $3n$, where n is the number expected. This number is then reduced to about n + 10% atoms by removing atoms in the lowest density. This procedure has the advantage that it places the free atoms at protein-like distances while covering the whole volume of the protein.

Such a free-atom model can describe almost every feature of an electron-density map, but this interpretation rarely provides a

Copyright © 2012 International Union of Crystallography

conventional picture of a protein. Nevertheless, information from parts of the improved map and the free-atom model can be used to recognize automatically real elements of protein structure by applying the model-reconstruction algorithms, so that at least a partial atomic protein model can be built. During iterative model building and refinement some free atoms gain chemical identity, whereas others remain free. Combination of this partial protein model with a free-atom set (a *hybrid* model) allows a considerably better description of the current map. The protein model provides additional information (in the form of stereochemical restraints), while prominent features in the electron density (unaccounted for by the current model) are described by free atoms. Importantly, use of the atomic positions in the hybrid model, rather than map grid points, as guides for model building in the electron-density maps allows the implementation of computationally more efficient algorithms.

18.8.3. *ARP/wARP* applications

Although *ARP/wARP* started life as a tool to build proteins, it now provides a wider spectrum of functionalities. As of the software version 7.1 (2009) *ARP/wARP* tackles several tasks in macromolecular crystallography:

(i) iterative protein-model building including a decision-making control module;
(ii) fast construction of the secondary structure of a protein;
(iii) sequence docking and side-chain building and refinement;
(iv) building flexible loops in alternate conformations;
(v) building DNA or RNA polynucleotide fragments;
(vi) fully automated placement of ligands, including a choice of the best-fitting ligand from a 'cocktail'; and
(vii) locating ordered water molecules.

ARP/wARP protocols are computationally efficient so that the time required is typically of the order of a few minutes on modern workstations, although iterative protein-model building may take a few hours. *ARP/wARP* applications are briefly described below, but for a detailed overview a reader is referred to Langer *et al.* (2008) or the literature cited hereafter.

18.8.3.1. Iterative protein-model building

If initial experimental phase information is available, the hybrid models are used as the main tool for obtaining as full a protein model as possible from the map calculated with the initial phases. A molecular-replacement solution can be directly used as an initial hybrid model. Given the information contained in the hybrid model in the traditional form of stereochemical restraints, refinement can work more efficiently, new improved phases can be obtained and a more accurate and complete protein model can be constructed. The new hybrid model is re-input to the refinement and these steps are iterated so that improved phases result in the construction of longer fragments of the protein chains. An almost complete protein model can often be obtained in a fully automated way (Fig. 18.8.3.1).

Automatic density-map interpretation for proteins is based on the location of the atoms in the current model and consists of several steps. Firstly, each pair of atoms in the free-atom model located at a distance of 3.8 ± 1.0 Å is assigned a potential peptide connection. The method utilizes the fact that all residues that comprise a protein, with the exception of *cis*-peptides, have chemically and structurally identical main-chain fragments with a highly characteristic, close-to-planar shape: the Cα—C—O—N—Cα *trans*-peptide units.

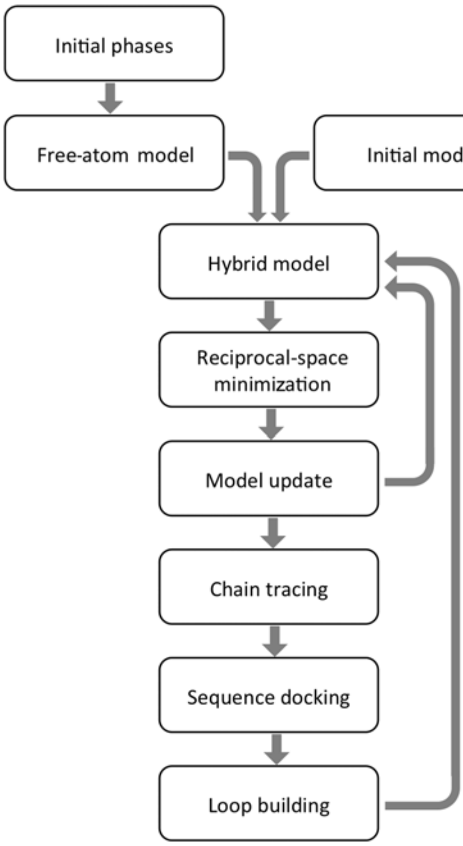


Figure 18.8.3.1
A flow chart of the iterative building of a protein structure with *ARP/wARP*.

The problem of searching for possible peptide units and their connections thus becomes straightforward. Peptide recognition between potential Cα pairs takes place by matching the electron density that surrounds each potential pair (Lamzin & Wilson, 1997) to that pre-computed for true Cα pairs from known structures. The peptide search uses density shapes sampled at 915 points, which account not only for the resolution of the X-ray data, but also for the fall-off in the diffraction intensity with resolution (the Wilson *B* factor). The important constraint here is that proteins are composed of linear non-branching polypeptide chains, allowing sets of connected peptides to be obtained from an initial list of all possible tracings. A limited depth-first graph-search algorithm is used for the assembly of peptides into linear polypeptide chain fragments for protein chain tracing (Morris *et al.*, 2002). Selection of the direction of a chain path is based on the electron density and observed backbone conformations. However, the set of peptide units and the list of how they are interconnected do not allow unambiguous tracing of a full-length chain for most structures.

The crucial bottleneck in automatically reconstructing a protein model from electron-density maps is achieving an initial trace of the polypeptide chain, typically approximated by many or hopefully only a few polypeptide fragments. Iteratively, these fragments should be made longer, their number should decrease and possible tracing errors, arising from poor starting phases, are eliminated. Following that, the fragments have to be placed in sequence, side chains built and gaps between the fragments that have been docked in sequence must be filled.

The complexity of the chain tracing can be illustrated as follows. Let us assume an example with 10 000 trial peptide units, where each free atom has, on average, two incoming and two outgoing peptides; this is typical for a good electron-density map of a 100 kDa (1000 residue) protein. Assuming that both ends of the chain are known and that it is possible to connect all the points regardless of the chosen route, then one is faced with the problem of choosing the best chain out of 2^{999} candidates. Screening all possible pairs of chain ends increases this number even further. Clearly, an exhaustive search is not computationally feasible. In practice, the situation is even more complex, since not all trial peptides are necessarily correctly identified in the first iteration, and some may be missing – analogous to the correctness or incorrectness of the atomic positions mentioned above. *ARP/wARP* uses a number of algorithmic shortcuts and condenses a vast amount of information, *e.g.* for the above example from 10^6 grid points of the map, through 10^4 free atoms, 2×10^4 peptide candidates, 5×10^4 dipeptide connections, down to 2×10^3 possible peptides and finally into a model with 10^4 atoms. Protein residues are differentiated only as glycine, alanine, serine and valine, and complete side chains are not built at this stage.

For every residue in the traced polypeptide fragments, a side-chain type is assigned with an empirical probability, using either connectivity criteria between free atoms and the α-carbon position, or rotamer conformations and their correlation to the density map (Cohen *et al.*, 2004). Given these guesses for the side chains and the known primary sequence, each polypeptide fragment is docked into sequence. Following this assignment, side chains are built in the best rotamer configuration.

After sequence docking, the missing parts of the model, typically extended and poorly ordered surface loops, can be identified. Using a distribution of five-Cα fragments derived from known structures, a number of structurally possible conformations are constructed. The procedure follows a pattern-recognition-based approach, where potential Cα candidates are looked for in non-negative density, followed by the construction of peptide planes and validation of dipeptide geometry using the Ramachandran plot. The loops are then scored based on the density and the best fitting ones are selected (Joosten *et al.*, 2008).

In essence, *ARP/wARP* improves the convergence properties of map interpretation during the course of protein building and refinement in the form of partial stereochemical assignments of the model. The landscape of the objective function used in refinement thus undergoes topological changes, which minimize model bias and favour convergence to a global minimum. Taken together, the probabilistic identification of the peptide units, the naturally high conformational flexibility of the connections of the peptide units, and the limited quality of the X-ray data and/or phases introduce large enough errors to cause density breaks in the middle of the chains or result in density overlaps. The result of such a tracing is usually a set of several main-chain fragments. The less accurate the starting maps (*i.e.* initial phases) and the lower the resolution and quality of the X-ray data, the more breaks there will be in the tracing and the higher the number of peptide units that will be difficult to identify.

18.8.3.2. Recognition of secondary structural elements

At a resolution of 3.0 Å and worse, where electron-density maps lack atomic features, *ARP/wARP* uses a different algorithm to build an initial model. Instead of free atoms, sparse map grid points with ~1 Å spacing are selected as potential Cα atoms on the basis of their density. A statistical pattern-classification approach based on supervised learning is applied. Numerical features corresponding to the distances and angles within building units larger than single peptides that have well defined patterns in electron-density maps are exploited using a nonlinear discriminant as the classification engine. This parameterizes the distribution of the feature values by a suitable hyperbody (*e.g.* an ellipsoid) and takes the body isosurfaces as decision boundaries. The discriminant threshold is set so as to minimize the number of false negatives in the net classification result. The process scores and filters fragments, and then assembles longer fragments iteratively. These undergo the same filtering with an adapted discriminant function, in a hierarchical order to make this procedure computationally inexpensive. The result is a set of Cα fragments that conform to known α-helical or β-stranded conformations. Finally, peptide backbone and Cβ atoms are added, the chain fragments are refined in real space, and the most probable chain direction is selected.

When the protein structure is (nearly) complete, smaller compounds (polynucleotides, ligands, cofactors and solvent) bound to the protein can be modelled in the difference electron-density map as outlined below.

18.8.3.3. Building polynucleotide fragments

Compared to proteins, crystals of nucleic acids in complex with proteins tend to have a less ordered crystalline arrangement, diffract less well and yield structures that exhibit higher atomic displacement parameters. Like the search for the secondary structural elements, *ARP/wARP* uses sparse grid points of the electron density and identifies candidate planar bases and phosphate groups that are expected to scatter strongly and be easier to discriminate (Hattne & Lamzin, 2008). The numerical algorithm is based on the use of third-order moment invariants. The method does not rely on any particular spatial arrangement of nucleotides and allows for wide structural variance in single- or double-stranded nucleic acids, especially in RNA. The problem of polynucleotide chain tracing is formulated in a manner analogous to that for proteins, but instead of peptides the search is carried out for phosphate–sugar–phosphate groups using pre-computed 915-point density shapes. The nucleotide chains are assembled like those of proteins with the difference that tracing a protein main chain accounts for two conformational degrees of freedom per residue, while a nucleotide unit has five.

18.8.3.4. Building bound ligands

To facilitate ligand building, all regions of difference density that have approximately the same volume as the ligand are identified. Subsequently, several numeric features of the density region and its sparse representation are used to produce an ensemble of putative ligand structures to best fit the local density. All ligand model candidates are refined in real space to fit the density shape and the single best model is chosen (Evrard *et al.*, 2007). A technique that compares the shapes of difference-electron-density blobs with the shape of the ligand to be built is used to distinguish compounds from a list (cocktail) of ligand candidates. The ligand that fits best is selected for further construction of the ensemble and subsequent restrained refinement.

18.8.3.5. Solvent building

In this application, the protein model (together with bound metals, ligands or nucleic acids) is not rebuilt during refinement. Instead, only the solvent structure is continuously updated. Since ordered solvent in a typical crystal structure comprises about

10% of the model, improvement of solvent indirectly improves the density corresponding to the protein part.

18.8.4. Iterations

Building protein chains or solvent with *ARP/wARP* proceeds in an iterative fashion. When the quality of the (partially built) model is sufficiently high, the phases improve overall and result in an enhanced electron density where a more accurate and more complete model may be built. An important component within iterations is the model update, described above, where parts of the existing model located in weak density are removed and new atoms added where the density acquires pronounced features. The iterative approach can also be extended to the building of ligands and nucleic acids.

18.8.5. Applicability and requirements

Model building with *ARP/wARP* is dependent on the resolution and quality of the X-ray data. Highly complete and accurate protein models are obtained at a resolution of 2.7 Å or higher provided that reasonable starting phase estimates are available. Success has been reported at lower resolution, but it is strongly dependent on the initial phase quality. Recent advances in refinement using experimental phasing information (Skubák *et al.*, 2009) proved essential for successful automated model building in a number of difficult test cases. The resolution of the X-ray data places similar limitations on *ARP/wARP* ligand building. Building secondary structural elements does not use free atoms and is applicable at a resolution of as low as 4.5 Å. We note that the quoted resolution limits relate to the current version of the software and are likely to change in the future.

References

Agarwal, R. C. & Isaacs, G. (1977). *Method for obtaining a high resolution protein map starting from a low resolution map. Proc. Natl Acad. Sci. USA*, **74**, 2835–2839.

Cohen, S. X., Morris, R. J., Fernandez, F. J., Ben Jelloul, M., Kakaris, M., Parthasarathy, V., Lamzin, V. S., Kleywegt, G. J. & Perrakis, A. (2004). *Towards complete validated models in the next generation of ARP/wARP. Acta Cryst.* D**60**, 2222–2229.

Evrard, G. X., Langer, G. G., Perrakis, A. & Lamzin, V. S. (2007). *Assessment of automatic ligand building in ARP/wARP. Acta Cryst.* D**63**, 108–117.

Hattne, J. & Lamzin, V. S. (2008). *Pattern-recognition-based detection of planar objects in three-dimensional electron-density maps. Acta Cryst.* D**64**, 834–842.

Joosten, K., Cohen, S. X., Emsley, P., Mooij, W., Lamzin, V. S. & Perrakis, A. (2008). *A knowledge-driven approach for crystallographic protein model completion. Acta Cryst.* D**64**, 416–424.

Lamzin, V. S. & Wilson, K. S. (1993). *Automated refinement of protein models. Acta Cryst.* D**49**, 129–147.

Lamzin, V. S. & Wilson, K. S. (1997). *Automated refinement for protein crystallography. Methods Enzymol.* **277**, 269–305.

Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. Nat. Protoc.* **3**, 1171–1179.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *ARP/wARP's model-building algorithms. I. The main chain. Acta Cryst.* D**58**, 968–975.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Automated protein model building combined with iterative structure refinement. Nat. Struct. Biol.* **6**, 458–463.

Skubák, P., Murshudov, G. & Pannu, N. S. (2009). *A multivariate likelihood SIRAS function for phasing and model refinement. Acta Cryst.* D**65**, 1051–1061.

International Tables for Crystallography (2012). Vol. F, Chapter 18.9, pp. 529–533.

Chapter 18.9. Macromolecular applications of *SHELX*

G. M. Sheldrick

18.9.1. Introduction

The first version of *SHELX* was written around 1970 for the solution and refinement of organic and inorganic structures. Since then, most routine structure determinations published in parts B, C and E of *Acta Crystallographica* have made use of at least one of the *SHELX* programs. However, this account will be restricted to the experimental phasing and refinement of macromolecular structures; details of other applications may be found in Sheldrick (2008). The *SHELX* programs are available as open source and as precompiled executables for common computer systems. They are written in a simple Fortran subset with a strict 'zero dependency' philosophy: the stand-alone binaries use no extra programs, libraries, data files or environmental variables *etc.*, which makes them easy to install and use. The programs are designed to be called from a command line and for incorporation into GUIs (graphical user interfaces); for example, experimental phasing with *SHELXC*, *SHELXD* and *SHELXE* may be performed *inter alia via* the *HKL2MAP* (Pape & Schneider, 2004), *CCP4i* (Potterton *et al.*, 2003), *AUTOSHARP* (Vonrhein *et al.*, 2007), *HKL-3000* (Minor *et al.*, 2006) and *CRANK* (Ness *et al.*, 2004) GUIs and the *AUTORICKSHAW* (Panjikar *et al.*, 2005) and *ARCIMBOLDO* (Rodriguez *et al.*, 2009) servers. The *SHELX* programs are general for all space groups in conventional settings or otherwise and make extensive use of default settings to keep user input and confusion to a minimum. Particular care has been taken to test the programs thoroughly on as many computer systems and crystallographic problems as possible before they were released, a process that often required several years!

18.9.2. Experimental phasing with *SHELXC/D/E*

Experimental phasing of macromolecules usually requires the presence of a relatively small number of atoms with significantly greater real or anomalous scattering power than the majority of atoms in the structure. These atoms can be *e.g.* metal or sulfur atoms already present in a native protein, heavy metals or halides introduced by soaking, or selenium incorporated by replacing methionine by selenomethionine using a suitable expression system. It is first necessary to locate the positions of these atoms in the unit cell, often using direct or Patterson methods and programs originally designed to solve small-molecule structures. These atoms (referred to as the substructure) are used to estimate starting phases for the complete macromolecule, which are then refined by modifying the electron density so that it corresponds better to the expected density whilst remaining consistent with the experimental native intensity data. In the *SHELX* system (Sheldrick, 2008, 2010), the program *SHELXC* is used to analyse and process the reflection data, and write the files that are then read by *SHELXD* for location of the substructure and *SHELXE* for derivation of the starting phases and density modification. In these programs, the experimental phasing is reduced to its absolute essentials, with the aim of obtaining an interpretable electron-density map quickly and reliably, rather than finding the most accurate phases. This requires some severe simplifications, for example the assumption that there is only one type of substructure atom present, although in practice a mixture of elements rarely causes problems. The approach has the advantage of producing robust, fast and simple to use programs but is restricted to experimental phasing by MAD (multi-wavelength anomalous dispersion), SAD (single-wavelength anomalous dispersion), SIR (single isomorphous replacement), SIRAS (combined SAD and SIR) and RIP (phasing based on radiation-induced changes in the structure) methods.

18.9.2.1. Substructure location with SHELXD

The program *SHELXD* (Usón & Sheldrick, 1999; Schneider & Sheldrick, 2002; Sheldrick, 2008: see also Chapter 16.1 in this volume) was originally designed for the *ab initio* solution of macromolecular structures from atomic resolution native data alone, but has proved very useful for substructure solution at much lower resolution. The dual-space approach of *SHELXD* was inspired by the *Shake and Bake* philosophy of Miller *et al.* (1993, 1994) but differs in many details, in particular in the use it makes of the Patterson function, which proves very effective in substructure solution. An advantage of the Patterson function is that it provides a good noise filter for the ΔF or F_A data: negative regions of the Patterson function can simply be ignored. On the other hand, the direct-methods approach is efficient at handling a large number of sites, whereas the number of Patterson peaks to analyse increases with the square of the number of atoms. Thus, for reasons of efficiency, the Patterson function is employed at two stages in *SHELXD*: at the beginning to obtain pseudo-random starting atom positions that are consistent with the Patterson (usually referred to as Patterson seeding) and to calculate a triangular table in which the off-diagonal elements are a Patterson minimum function for vectors between one potential atom and a second atom including all its symmetry equivalents (Schneider & Sheldrick, 2002). This table, which also showed the minimum distances between atoms taking symmetry into account, was very useful for recognizing NCS (noncrystallographic symmetry) in substructures obtained from borderline data, but proved difficult to automate and so has remained a 'crossword table' for the amusement of experts (it is still written by *SHELXD* to the listing file). Similarly, a figure of merit PATFOM calculated by *SHELXD* on the basis of this table has proved less useful than expected in practice, because the values are not on an absolute scale; they tend to be much higher for small substructures.

18.9.2.2. Practical considerations for substructure solution

The resolution to which the data have to be truncated is usually the most critical decision for substructure solution. When the anomalous signal is weak, using data out to the diffraction limit introduces noise and may well not lead to a solution. On the other hand, truncating the resolution too strongly leads to less precise atomic positions and may result in weaker sites being missed. The mean ratio of the average absolute anomalous difference to its

Copyright © 2012 International Union of Crystallography

estimated standard deviation can be used as a guide as to where to truncate the data, but this requires good estimates of the standard deviations of the measured intensities. A better guide is the correlation coefficient between the signed anomalous differences for the different wavelengths in a MAD experiment (Schneider & Sheldrick, 2002), or for two different crystals in a SAD experiment. In a sulfur SAD experiment, the two sulfur atoms in a disulfide bridge behave as a single 'super-sulfur' atom at resolutions less than about 2.1 Å. Although this is no impediment to finding them with *SHELXD*, their phasing power is significantly enhanced if they are resolved into their two component sulfur atoms using the DSUL instruction in *SHELXD*.

To allow for possible variations in occupancy or displacement parameters (B values) and the presence of different types of marker atoms, it has proved useful to refine the occupancies in the last two dual-space cycles. A sharp fall-off in the refined occupancy between the last true site and the first noise peak is also a useful test for a good solution, but cannot be used for halide soaks, for which a continuous range of occupancies is usually found; usually the peak list is truncated at about 0.2 times the occupancy of the highest peak, because experience indicates that weaker peaks may well be noise. The correlation coefficient (CC) between the observed and calculated E values usually enables correct solutions to be identified unambiguously, and the value of CC(weak), the correlation coefficient based on the reflections not used in the dual-space recycling, is also a good check. It is like a free R value, but is not quite independent because all the data are used in the occupancy refinement.

For large substructures, it is important not to terminate the search after too small a number of trials. Often, a much better solution is found if more trials are performed. Several structures have given only one correct solution in 10 000 trials (and in one case only one in 1.6 million trials!). Since the default *SHELXD* action is to start each job with a different random number seed, one efficient approach is to run several copies in parallel, *e.g.* one per CPU (Fu *et al.*, 2007). This may conveniently be combined with different values for the resolution cutoff.

18.9.2.3. SHELXE: density modification

The density modification in *SHELXE* does not make use of solvent flattening or histogram matching. Instead, the *sphere of influence algorithm* (Sheldrick, 2002) is used to provide an indication as to how likely it is that each individual voxel (volume element) in the map corresponds to a true atomic site. The variance V of the density on a spherical surface of radius 2.42 Å is calculated for each voxel in the map. The use of a spherical surface rather than a spherical volume was intended to save time and add a little chemical information (2.42 Å is a typical 1,3 distance in proteins and DNA). V gives an indication of the probability that a voxel corresponds to a true atomic position. Voxels with low V are *flipped* ($\rho' = -\rho\gamma$ where γ is usually set to 1.0). For voxels with high V, ρ is replaced either by $\{\rho^4/[\upsilon^2\sigma^2(\rho) + \rho^2]\}^{1/2}$ [where ν is usually 0.5 and $\sigma^2(\rho)$ is the variance of the density ρ over the whole cell] if positive or by zero if negative. An empirical weighting scheme for phase recombination is used to combat model bias. The variance over all voxels in the asymmetric unit of the individual variances V, output by the program as the 'contrast', is a good indication of which marker atom enantiomorph is correct; it is almost invariably higher for the correct choice, especially after 5–10 density-modification cycles.

A further simple and effective algorithm to improve the phases of the experimentally measured reflections is to extrapolate the data and phases to a higher resolution than was actually accessible: the *free lunch algorithm* (Caliandro *et al.*, 2005; Yao *et al.*, 2006). This has also been implemented in *SHELXE* (Usón *et al.*, 2007). This algorithm is effective when data have been measured to a resolution of 2.0 Å or better, and can lead to improvements in the mean phase error of the measured reflections of between 5 and 30°.

18.9.2.4. Integrated density modification and autotracing

A relatively fast iterative autotracing algorithm has been incorporated into the density modification in *SHELXE* (Sheldrick, 2010). It is primarily designed to get a toehold in maps with very poor starting phases, *e.g.* with a mean phase error greater than 60°. The tracing proceeds as follows:

(i) Find potential α-helices in the density and try to extend them at both ends. Then find other potential tripeptides and try to extend them at both ends in the same way.
(ii) Add traces obtained by applying NCS operations derived from the heavy-atom positions, provided that they fit the density well.
(iii) Tidy up and splice the traces as required.
(iv) Use the traced residues to estimate phases and combine these with the initial phase information using σ_A weights (Read, 1986), then restart the density modification. The refinement of one B value per residue provides a further opportunity to suppress wrongly traced residues.

Crystallographic symmetry is taken into account at all stages, and the backbone traces are selected with the help of the following criteria:

(i) The modified density ρ' should be high at the atomic sites and low at the dummy-atom positions (generated at sites where no atom should be found).
(ii) The chains must be long enough (in general at least seven amino acids); longer chains are given a higher weight.
(iii) A few Ramachandran outliers can be tolerated, *e.g.* for glycines, but in general the φ and ψ angle pairs should lie in the well populated regions of the Ramachandran diagram.
(iv) There should be a well defined secondary structure (*i.e.* φ/ψ pairs should tend to be similar for consecutive residues).
(v) On average, there should be significant positive density 2.9 Å from N in the N→H direction (to a hydrogen-bond acceptor). This takes into account that the large majority of main-chain NH groups in proteins take part in hydrogen bonds with oxygen or other electronegative atoms.

At the time of writing, the autotracing version of *SHELXE* is still at the β-test stage, but the results are encouraging. The ability to trace the structure in a way that is consistent with the above criteria is a very good indicator as to whether the structure has been solved. Two internet servers for experimental phasing that use this version effectively are *AUTORICKSHAW* (Panjikar *et al.*, 2005) and *ARCIMBOLDO* (Rodríguez *et al.*, 2009). In the latter, Usón and co-workers use this ability to trace maps with large initial phase errors, and so to find a few correct solutions in a large number of trials generated by molecular-replacement searches for short α-helices, as part of a procedure for the *ab initio* solution of protein structures with data down to 2.0 Å resolution (see Chapter 16.1).

18.9.3. Macromolecular refinement using *SHELXL*

SHELXL is a very general refinement program and is equally suitable for the refinement of minerals, organometallic structures, oligonucleotides or proteins (or any mixture thereof) against X-ray or neutron single- (or twinned!) crystal data. It has even been used with diffraction data from powders, fibres and two-dimensional crystals. The price of this generality is that it is somewhat slower than programs specifically written only for protein structure refinement. Structure factors are calculated by summation rather than by a fast Fourier transform; this is more precise but slower. Any protein- (or DNA-)specific information must be input to *SHELXL* by the user in the form of refinement restraints *etc.* Refinement of macromolecules using *SHELXL* has been discussed by Sheldrick & Schneider (1997).

Least-squares (LS) refinement has the advantage that it provides proper estimates of the standard deviations in all refined and derived parameters. However, at low data-to-parameter ratios, maximum likelihood (ML) refinement is better at avoiding overfitting and is less sensitive to reflection outliers. For this reason, ML refinement, as implemented in, for example, *REFMAC* (Murshudov *et al.*, 1997), *PHENIX_REFINE* (Adams *et al.*, 2002) or *BUSTER-TNT* (Blanc *et al.*, 2004), is recommended when the data-to-parameter ratio falls below about 5. ML refinement also enables experimental phase information to be incorporated (Pannu *et al.*, 1998).

18.9.3.1. *Constraints and restraints*

In refining macromolecular structures, it is almost always necessary to supplement the diffraction data with chemical information in the form of *restraints*. A typical restraint is the condition that a bond length should approximate to a target value with a given estimated standard deviation; restraints are treated as extra experimental data items. Even if the crystal diffracts to 1.0 Å, there may well be poorly defined disordered regions for which restraints are essential to obtain a chemically sensible model (the same can be true of small molecules too!). *SHELXL* is generally not suitable for refinements at resolutions lower than about 2.0 Å because it cannot currently handle general potential-energy functions, *e.g.* for torsion angles or hydrogen bonds; if noncrystallographic symmetry restraints can be employed, this limit can be relaxed a little. Unimodal restraints, such as target distances, are less likely to result in local minima than are multimodal restraints, such as torsion angles; multimodal functions are better used as validation criteria. It is fortunate that validation programs, such as *MolProbity* (Davis *et al.*, 2007), make good use of multimodal functions such as torsion angles and hydrogen-bonding patterns that are not employed as restraints in *SHELXL* refinements.

For some purposes (*e.g.* riding hydrogen atoms, rigid-group refinement, or occupancies of atoms in disordered side chains), *constraints*, exact conditions that lead to a reduction in the number of variable parameters, may be more appropriate than restraints; *SHELXL* allows such constraints and restraints to be mixed freely. Riding hydrogen atoms are defined such that the C—H vector remains constant in magnitude and direction, but the carbon atom is free to move; the same shifts are applied to both atoms, and both atoms contribute to the least-squares derivative sums. This model may be combined with anti-bumping restraints that involve hydrogen atoms, which helps to avoid unfavourable side-chain conformations. *SHELXL* also provides, *e.g.*, methyl groups that can rotate about their local threefold axes; the initial torsion angle may be found using a difference-electron-density synthesis calculated around the circle of possible hydrogen-atom positions.

18.9.3.2. *Least-squares refinement algebra*

The original *SHELX* refinement algorithms were modelled closely on those described by Cruickshank (1970). For macromolecular refinement, an alternative to (blocked) full-matrix refinement is provided by the conjugate-gradient solution of the least-squares normal equations as described by Hendrickson & Konnert (1980), including preconditioning of the normal matrix that enables positional and displacement parameters to be refined in the same cycle. The structure-factor derivatives contribute only to the diagonal elements of the normal matrix, but all restraints contribute fully to both the diagonal and non-diagonal elements, although neither the Jacobian nor the normal matrix itself are ever generated by *SHELXL*. The parameter shifts are modified by comparison with those in the previous cycle to accelerate convergence whilst reducing oscillations. Thus, a larger shift is applied to a parameter when the current shift is similar to the previous shift, and a smaller shift is applied when the current and previous shifts have opposite signs. The STIR instruction enables the resolution to be extended stepwise during the refinement, thereby increasing the radius of convergence, and is recommended for the first refinement job starting from an unrefined model. It can also be used to reduce memory effects if the $R_{\rm free}$ reflection set has been modified.

SHELXL refines against F^2 rather than F, which enables all data to be used in the refinement with weights that include contributions from the experimental uncertainties, rather than having to reject F values below a preset threshold; there is a choice of appropriate weighting schemes. Provided that reasonable estimates of $\sigma(F^2)$ are available, this enables more experimental information to be employed in the refinement; it also allows refinement against data from twinned crystals.

18.9.3.3. *Full-matrix estimates of standard uncertainties*

Inversion of the full normal matrix (or of large matrix blocks, *e.g.* for all positional parameters) enables the precision of individual parameters to be estimated (Rollett, 1970), either with or without the inclusion of the restraints in the matrix. The standard uncertainties in dependent quantities (*e.g.* torsion angles or distances from mean planes) are calculated in *SHELXL* using the full least-squares correlation matrix. These standard uncertainties reflect the data-to-parameter ratio, *i.e.* the resolution and completeness of the data and the percentage of solvent, and the quality of the agreement between the observed and calculated F^2 values (and the agreement of restrained quantities with their target values when restraints are included).

Full-matrix refinement is also useful (together with the stepwise increase of resolution STIR) when domains are refined as rigid groups in the early stages of refinement (*e.g.* after structure solution by molecular replacement), since the total number of parameters is small and the correlation between parameters may be large.

18.9.3.4. *Refinement of anisotropic displacement parameters*

The motion of macromolecules is clearly anisotropic, but the data-to-parameter ratio rarely permits the refinement of the six independent anisotropic displacement parameters (ADPs) per atom; even for small molecules and data to atomic resolution, the anisotropic refinement of disordered regions requires the use of restraints. *SHELXL* employs three types of ADP restraint

(Sheldrick 1993; Sheldrick & Schneider, 1997). The *rigid bond restraint*, first suggested by Rollett (1970), assumes that the components of the ADPs of two atoms connected *via* one (or two) chemical bonds are equal within a specified standard deviation. This has been shown to hold accurately (Hirshfeld, 1976; Trueblood & Dunitz, 1983) for precise structures of small molecules, so it can be applied as a *hard* restraint with small estimated standard deviation. The *similar ADP restraint* assumes that atoms that are spatially close (but not necessarily bonded, because they may be different components of a disordered group) have similar U^{ij} components. An *approximately isotropic restraint* is useful for isolated solvent molecules. These two restraints are only approximate and so should be applied with low weights, *i.e.* high estimated standard deviations.

The transition from isotropic to anisotropic roughly doubles the number of parameters and almost always results in an appreciable reduction in the R factor. However, this represents an improvement in the model only when it is accompanied by a significant reduction in the free R factor (Brünger, 1992). Since the free R factor is itself subject to uncertainty because of the small sample used, a drop of at least 1% is needed to justify anisotropic refinement. The resulting displacement ellipsoids should make chemical sense and not be '*non-positive-definite*'!

18.9.3.5. Similar geometry and NCS restraints

When there are several identical chemical moieties in the asymmetric unit, a very effective restraint is to assume that the chemically equivalent 1,2 and 1,3 distances are the same, but unknown. This technique is easy to apply using *SHELXL* and is often employed for small-molecule structures and, in particular, for oligosaccharides. Similarly, the P—O bond lengths and six O···O distances in all phosphate anions can be assumed to be the same (but without a target value), *i.e.* it is assumed that the whole crystal is at the same pH. Local noncrystallographic symmetry (NCS) restraints (Usón *et al.*, 1999) may be applied to restrain corresponding 1,4 distances and isotropic displacement parameters to be the same when there are several identical macromolecular domains in the asymmetric unit; usually, the 1,2 and 1,3 distances are restrained to standard values in such cases and so do not require NCS restraints. Such *local NCS restraints* are more flexible than *global NCS constraints* and – unlike the latter – do not require the specification of a transformation matrix and mask.

18.9.3.6. Modelling disorder and solvent

There are many ways of modelling disorder using *SHELXL*, but for macromolecules the most convenient is to retain the same atom and residue names for the two or more components and assign a different 'part number' (analogous to the PDB alternative site flag) to each component. With this technique, no change is required to the input restraints *etc.* Atoms in the same component will normally have a common occupancy that is assigned to a 'free variable'. If there are only two components, the sum of their occupancies can be constrained to be unity; if there are more than two components, the sum of their free variables may be restrained to be unity. Since any linear restraint may be applied to the free variables, they are very flexible, *e.g.* for modelling complicated disorder. By restraining distances to be equal to a free variable, a standard deviation of the mean distance may be calculated rigorously using full-matrix least-squares algebra.

Babinet's principle is used to define a bulk solvent model with two refinable parameters (Moews & Kretsinger, 1975), and global anisotropic scaling (Usón *et al.*, 1999) may be applied using a parameterization proposed by Parkin *et al.* (1995). An auxiliary program, *SHELXWAT*, allows automatic water divining by iterative least-squares refinement, rejection of waters with high displacement parameters, difference-electron-density calculation, and a peak search for potential water molecules that make at least one good hydrogen bond and no bad contacts; this is a simplified version of the *ARP* procedure of Lamzin & Wilson (1993).

18.9.3.7. Twinned crystals

SHELXL provides facilities for refining against data from merohedral, pseudo-merohedral and non-merohedral twins (Herbst-Irmer & Sheldrick, 1998). Refinement against data from merohedrally twinned crystals is particularly straightforward, requiring only the twin law (a 3 × 3 matrix) and starting values for the volume fractions of the twin components. Failure to recognize such twinning not only results in high R factors and poor quality maps, it can also lead to incorrect biochemical conclusions (Luecke *et al.*, 1998). Twinning can often be detected by statistical tests (Yeates & Fam, 1999), and it is probably much more widespread in macromolecular crystals than is generally appreciated!

18.9.4. *SHELXPRO* – protein interface to *SHELX*

The *SHELX* system includes several auxiliary programs, the most important of which for macromolecular users is *SHELXPRO*. *SHELXPRO* provides an interface between *SHELXL* and other programs commonly used by protein crystallographers. Often, *SHELXL* will be used only for the final stages of refinement, in which case *SHELXPRO* is used to generate the *name.ins* file from a PDB format file, inserting the necessary restraints and other instructions. A corresponding .hkl file may be generated using the programs *XPREP* (Bruker AXS, Madison, Wisconsin) or *MTZ2HKL* (Grune, 2008). *SHELXPRO* is also used to prepare the *name.ins* file for a new refinement job based on the results of the previous refinement and to prepare data for PDB deposition. In addition, the refinement results can be summarized graphically in the form of PostScript plots.

18.9.5. Distribution and support of *SHELX*

The *SHELX* system is available free to academics and, for a small licence fee, to commercial users. The programs are supplied as Fortran sources and as precompiled versions for widely used operating systems. The programs, examples and extensive documentation may be downloaded from the Internet or (if necessary) supplied on CD ROM. Details of new developments, answers to frequently asked questions, and information about obtaining and installing the programs are available from the SHELX home page, http://shelx.uni-ac.gwdg.de/SHELX/. The author is always interested to receive reports of problems and suggestions for improving the programs and their documentation by e-mail (gsheldr@shelx.uni-ac.gwdg.de).

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *PHENIX: building new software for* *automated crystallographic structure determination*. *Acta Cryst.* D**58**, 1948–1954.

Blanc, E., Roversi, P., Vonrhein, C., Flensburg, C., Lea, S. M. & Bricogne, G. (2004). *Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT*. *Acta Cryst.* D**60**, 2210–2221.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005). *Phasing at resolution higher than the experimental resolution*. *Acta Cryst.* D**61**, 556–565.

Cruickshank, D. W. J. (1970). *Least-squares refinement of atomic parameters*. In *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 187–197. Copenhagen: Munksgaard.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids*. *Nucleic Acid Res.* **35**, W375–W383.

Fu, Z.-Q., Chrzas, J., Sheldrick, G. M., Rose, J. & Wang, B.-C. (2007). *A parallel program using SHELXD for quick heavy-atom partial structure solution on high-performance computers. J. Appl. Cryst.* **40**, 387–390.

Grune, T. (2008). *mtz2sca and mtz2hkl: facilitated transition from CCP4 to the SHELX program suite*. *J. Appl. Cryst.* **41**, 217–218.

Hendrickson, W. A. & Konnert, J. H. (1980). *Incorporation of stereochemical information into crystallographic refinement*. In *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.23. Bangalore: Indian Academy of Sciences.

Herbst-Irmer, R. & Sheldrick, G. M. (1998). *Refinement of twinned structures with SHELXL97. Acta Cryst.* B**54**, 443–449.

Hirshfeld, F. L. (1976). *Can X-ray data distinguish bonding effects from vibrational smearing? Acta Cryst.* A**32**, 239–244.

Lamzin, V. S. & Wilson, K. S. (1993). *Automated refinement of protein models. Acta Cryst.* D**49**, 129–147.

Luecke, H., Richter, H. T. & Lanyi, J. K. (1998). *Proton transfer pathways in bacteriorhodopsin at 2.3 Å resolution. Science*, **280**, 1934–1937.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *On the application of the minimal principle to solve unknown structures. Science*, **259**, 1430–1433.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB: crystal structure determination via Shake-and-Bake. J. Appl. Cryst.* **27**, 613–621.

Minor, W., Cymborowski, M., Otwinowski, Z. & Chruszcz, M. (2006). *HKL-3000: the integration of data reduction and structure solution – from diffraction images to an initial model in minutes*. *Acta Cryst.* D**62**, 859–866.

Moews, P. C. & Kretsinger, R. H. (1975). *Refinement of carp muscle parvalbumin by model building and difference Fourier analysis. J. Mol. Biol.* **91**, 201–228.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Ness, S. R., de Graaff, R. A. G., Abrahams, J. P. & Pannu, N. S. (2004). *Crank: New methods for automated macromolecular crystal structure solution. Structure*, **12**, 1753–1761.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment*. *Acta Cryst.* D**61**, 449–457.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst.* D**54**, 1285–1294.

Pape, T. & Schneider, T. R. (2004). *HKL2MAP: a graphical user interface for macromolecular phasing with SHELX programs*. *J. Appl. Cryst.* **37**, 843–844.

Parkin, S., Moezzi, B. & Hope, H. (1995). *XABS2: an empirical absorption correction program. J. Appl. Cryst.* **28**, 53–56.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *A graphical user interface to the CCP4 program suite*. *Acta Cryst.* D**59**, 1131–1137.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Rodriguez, D. D., Grosse, C., Himmel, S., Gonzales, C., de Ilarduya, I. M., Sheldrick, G. M. & Usón, I. (2009). *Crystallographic ab initio protein structure solution below atomic resolution*. *Nat. Methods*, **6**, 651–653.

Rollett, J. S. (1970). *Least-squares procedures in crystal structure analysis*. In *Crystallographic Computing*, edited by F. R. Ahmed, S. R. Hall & C. P. Huber, pp. 167–181. Copenhagen: Munksgaard.

Schneider, T. R. & Sheldrick, G. M. (2002). *Substructure solution with SHELXD*. *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (1993). *Refinement of large small-molecule structures using SHELXL-92*. In *Crystallographic Computing 6. A Window on Modern Crystallography*, edited by H. D. Flack, L. Párkányi & K. Simon, pp. 111–122. Oxford: IUCr and Oxford University Press.

Sheldrick, G. M. (2002). *Macromolecular phasing with SHELXE*. *Z. Kristallogr.* **217**, 644–650.

Sheldrick, G. M. (2008). *A short history of SHELX*. *Acta Cryst.* A**64**, 112–122.

Sheldrick, G. M. (2010). *Experimental phasing with SHELXC/D/E: combining chain tracing with density modification. Acta Cryst.* D**66**, 479–485.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high resolution refinement. Methods Enzymol.* **277**, 319–343.

Trueblood, K. N. & Dunitz, J. D. (1983). *Internal molecular motions in crystals. The estimation of force constants, frequencies and barriers from diffraction data. A feasibility study. Acta Cryst.* B**39**, 120–133.

Usón, I., Pohl, E., Schneider, T. R., Dauter, Z., Schmidt, A., Fritz, H.-J. & Sheldrick, G. M. (1999). *1.7 Å structure of the stabilised REI*$_v$ *mutant T39K. Application of local NCS restraints. Acta Cryst.* D**55**, 1158–1167.

Usón, I. & Sheldrick, G. M. (1999). *Advances in direct methods for protein crystallography*. *Curr. Opin. Struct. Biol.* **9**, 643–648.

Usón, I., Stevenson, C. E. M., Lawson, D. M. & Sheldrick, G. M. (2007). *Structure determination of the O-methyltransferase NovP using the 'free lunch algorithm' as implemented in SHELXE*. *Acta Cryst.* D**63**, 1069–1074.

Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). *Automated structure solution with autoSHARP. Methods Mol. Biol.* **364**, 215–230.

Yao, J. X., Dodson, E. J., Wilson, K. S. & Woolfson, M. M. (2006). *ACORN: a review*. *Acta Cryst.* D**62**, 901–908.

Yeates, T. O. & Fam, B. C. (1999). *Protein crystals and their evil twins. Structure*, **7**, R25–R29.

International Tables for Crystallography (2012). Vol. F, Chapter 18.10, pp. 534–538.

Chapter 18.10. *PrimeX* and the Schrödinger computational chemistry suite of programs

J. A. Bell, Y. Cao, J. R. Gunn, T. Day, E. Gallicchio, Z. Zhou, R. Levy and R. Farid

18.10.1. Introduction

PrimeX is a new crystal structure refinement program for large biological molecules. Among the tools included in the package are reciprocal- and real-space minimization, automated construction of polypeptides into electron density, side-chain placement, ligand placement and simulated-annealing refinement. *PrimeX* is unique among X-ray refinement packages for large biological molecules with respect to its origins: the popular docking program, *Glide*, and the protein structure prediction and refinement package, *Prime*.

18.10.1.1. Computational environment

PrimeX is fully integrated with *Maestro*, a general-purpose molecular graphics front end. It shares a consistently designed interface with a suite of other powerful programs in molecular modelling and computational chemistry. Some of these other programs and their functions are listed in Table 18.10.1.1.

18.10.1.2. Mission

The first goal of *PrimeX* is to produce protein models consistent with both the X-ray diffraction data and the needs of computational chemists. Each of the software packages listed in Table 18.10.1.1 depends, to some degree, on accurate protein structures as input for computations. However, the models of biological molecules produced by crystallographers typically cause problems for computational work in two ways. First, these models lack coordinates for hydrogen atoms. Because the experimental data rarely give any direct indication of the correct position of hydrogen atoms, crystallographers are reluctant to include them in their models. Secondly, these models frequently contain coordinates that place non-bonded atoms implausibly close together, leading to the calculation of extremely high internal energy in these molecules. Crystallographers historically have not restrained these interactions as strongly as they restrain other elements of molecular geometry such as bond lengths, bond angles and group planarity. A reason for placing atoms in these conflicting positions is to fit the atoms at the centre of the observed electron-density peak, even if, because of atomic motion, that peak extends over a large volume of space that intersects with the volume occupied by other atoms. The result is a model that may be representative of the time-averaged structure but which cannot be consistent with the structure at a given instance. *PrimeX* produces a model that is consistent with both timescale views.

Table 18.10.1.1
Schrödinger software packages related to *PrimeX*

Package	Brief description
CombiGlide	Lead compound design and optimization
Desmond	High-performance molecular dynamics simulations
Glide	Ligand–receptor docking and scoring
Induced-Fit Docking	Modelling conformational changes induced by ligand binding
Liason	Ligand–receptor binding free-energy prediction
MacroModel	Full-featured program for molecular modelling
Prime	Protein structure prediction and refinement
QSite	Quantum mechanics/molecular mechanics calculations on biological systems
SiteMap	Binding-site identification and analysis

The second goal of *PrimeX* is to minimize unnecessary experimenter interaction for repetitive tasks. At its current stage of development, *PrimeX* provides a comprehensive set of features for automation of the refinement of the X-ray crystal structures of large biological molecules. Future development will combine these tools in automated multi-step processes.

18.10.1.3. Implementation path

In a field where most of the available software has been developed from academic sources, *PrimeX* is the only refinement program for biological crystallography that has been developed in a commercial environment. The advantage of commercial development includes high standards for consistency, professional software support and specific attention to the concerns of crystallographers in industry. All code is developed internally and intellectual property rights are monitored carefully. In an industrial setting, this assurance may be important to avoid future intellectual property entanglements.

18.10.2. Computational environment

Several aspects of the software packages listed in Table 18.10.1.1 are especially relevant to the application of *PrimeX*.

18.10.2.1. Molecular graphics

Maestro is the unified interface for all Schrödinger software. It has impressive rendering capabilities, a powerful selection of analysis tools and a set of fundamental tools for manipulating molecules. Hardware stereo is supported. Alternate conformations are displayed and can be manipulated in *Maestro*. Automation of complex tasks and development of custom applications can be accomplished through integrated Python scripting in *Maestro*, as well as through an internal macro facility. Function keys are also customizable. The results of each operation in a series of steps are organized in a project table (Fig. 18.10.2.1). This feature of *Maestro* is well suited to documenting the multiple steps typical of crystal structure refinement.

18.10.2.2. Supporting infrastructure

All Schrödinger programs are accessed through a clear, well organized graphical user interface (GUI). Nearly all computations are run through Schrödinger's job control facility. This facility provides a uniform mechanism for launching, monitoring and controlling calculations. Jobs can be submitted *from* any computer *to* any computer, cluster or grid, or to a batch queue, and the results returned to the computer from which the job was submitted. In addition, many of Schrödinger's tools have been integrated as *KNIME* extensions. *KNIME* (http://www.knime.org) is an open-source pipelining program which provides a graphical interface for building workflows. *PrimeX* components will also be available for *KNIME* workflows in the future.

Copyright © 2012 International Union of Crystallography

18.10.2.3. *Molecular mechanics*

OPLS (Jorgensen *et al.*, 1996; Kaminski *et al.*, 2001) is a general-purpose force field for modelling proteins, nucleic acids and small molecules. This accurate description of atomic level interactions provides consistency across all Schrödinger molecular modelling packages, such as the core modelling functions of the *MacroModel* package or the molecular dynamics simulation tools of *Desmond*. *PrimeX* applies this consistent molecular description to the refinement of large biological crystal structures.

18.10.3. Achieving the mission of *PrimeX*

The main goal of any X-ray crystal structure refinement program is to produce an optimized fit to the X-ray diffraction data, consistent with known molecular stereochemistry. For *PrimeX*, the additional goal is to include hydrogen atoms in the model in a way that correctly reflects molecular stereochemistry, especially for non-bonded interactions.

18.10.3.1. *Molecular models and computational chemistry*

18.10.3.1.1. *Hydrogen atoms*

Although crystallographers have generally not included coordinates for hydrogen atoms in their models at moderate resolution, this situation has not always been so. During the early years of development of the very popular *XPLOR* package for refinement (Brünger, 1992), which was later augmented to become the even more popular refinement package *CNS* (Brünger *et al.*, 1998), the geometric restraints were based on the CHARMm force field. The presence of all polar hydrogen atoms, *i.e.* hydrogen atoms attached to non-carbon atoms, was required by this force field. Quite a number of structures from that era remain in the Protein Data Bank and include coordinates for polar hydrogen atoms. This CHARMm-based restraint system was later modified to a unified-atom model that is used by *CNS* today.

Use of the all-atom force field OPLS in *PrimeX* requires the explicit definition of all hydrogen-atom positions for geometric calculations. Hydrogen atoms are added to the model automatically in a manner that satisfies local hydrogen-bonding opportunities with adjacent atoms, if they are not already present from previous cycles of refinement. An additional hydrogen-bond optimization tool in *PrimeX* (Fig. 18.10.3.1) analyses donors and acceptors to define clusters of such sites that might be connected through hydrogen bonds, and that form networks. Within each cluster, hydrogen bonding is evaluated using a rule-based Monte Carlo method to find the optimal combinations of the variable components of these systems. The structural features adjusted during hydrogen-bond optimization are:

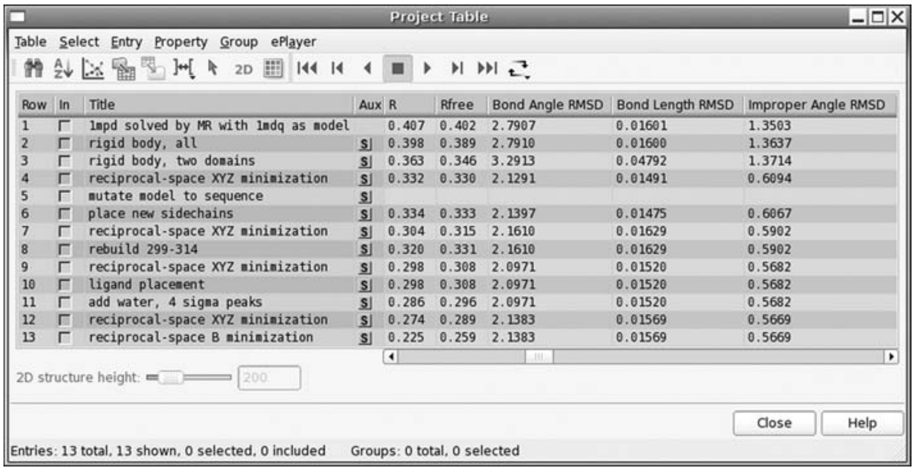
(i) alcoholic hydrogen-atom positions;
(ii) sulfhydril hydrogen-atom positions;
(iii) phenolic hydrogen-atom positions;
(iv) charge and tautomeric states of Asp and Glu side chains;
(v) charge, tautomeric states and orientation (flip) of His side chains;
(vi) orientation (flip) of Asn and Gln side chains; and
(vii) positions for hydrogen atoms in water molecules.

18.10.3.1.2. *Electrostatics and implicit solvation*

For hydrogen atoms bonded to more electronegative atoms, electrostatic forces are a major determinant of non-bonded interactions, and they must be evaluated during the calculation of geometric gradients in the refinement. Thus, *PrimeX* employs the complete molecular mechanics description of atomic interactions embodied in OPLS, including electrostatic terms. A solvation energy term (Zhu, Shirts & Friesner, 2007) can account for the effects of solvation on electrostatic interactions without the use of explicit coordinates for all solvent atoms. This term may be optionally included in the molecular-mechanics description used by *PrimeX*. Refinement incorporating a complete electrostatic description, including an implicit solvation term, has been shown to lead to lower $R_{\rm free}$ values compared with refinement excluding those interactions (Knight *et al.*, 2008). For cases where coordinates for bound water molecules were explicitly included in the refinement, the effects of an implicit solvation term on $R_{\rm free}$ were favourable but small (J. A. Bell, unpublished work). A more complete analysis of the effects of the implicit solvation term in *PrimeX* is in progress.

18.10.3.1.3. *van der Waals interactions*

The full OPLS force field includes terms for repulsive and attractive non-bonded interactions. Most significant for restraints during crystallographic refinement is the utilization of the full repulsive term for non-bonded interactions, unlike other leading biological crystal structure refinement packages where these terms are weighted relatively lightly.



Row	In	Title	Aux	R	Rfree	Bond Angle RMSD	Bond Length RMSD	Improper Angle RMSD
1	☐	1mpd solved by MR with 1mdq as model		0.407	0.402	2.7907	0.01601	1.3503
2	☐	rigid body, all	S	0.398	0.389	2.7910	0.01600	1.3637
3	☐	rigid body, two domains	S	0.363	0.346	3.2913	0.04792	1.3714
4	☐	reciprocal-space XYZ minimization	S	0.332	0.330	2.1291	0.01491	0.6094
5	☐	mutate model to sequence	S					
6	☐	place new sidechains	S	0.334	0.333	2.1397	0.01475	0.6067
7	☐	reciprocal-space XYZ minimization	S	0.304	0.315	2.1610	0.01629	0.5902
8	☐	rebuild 299-314	S	0.320	0.331	2.1610	0.01629	0.5902
9	☐	reciprocal-space XYZ minimization	S	0.298	0.308	2.0971	0.01520	0.5682
10	☐	ligand placement	S	0.298	0.308	2.0971	0.01520	0.5682
11	☐	add water, 4 sigma peaks	S	0.286	0.296	2.0971	0.01520	0.5682
12	☐	reciprocal-space XYZ minimization	S	0.274	0.289	2.1383	0.01569	0.5669
13	☐	reciprocal-space B minimization	S	0.225	0.259	2.1383	0.01569	0.5669

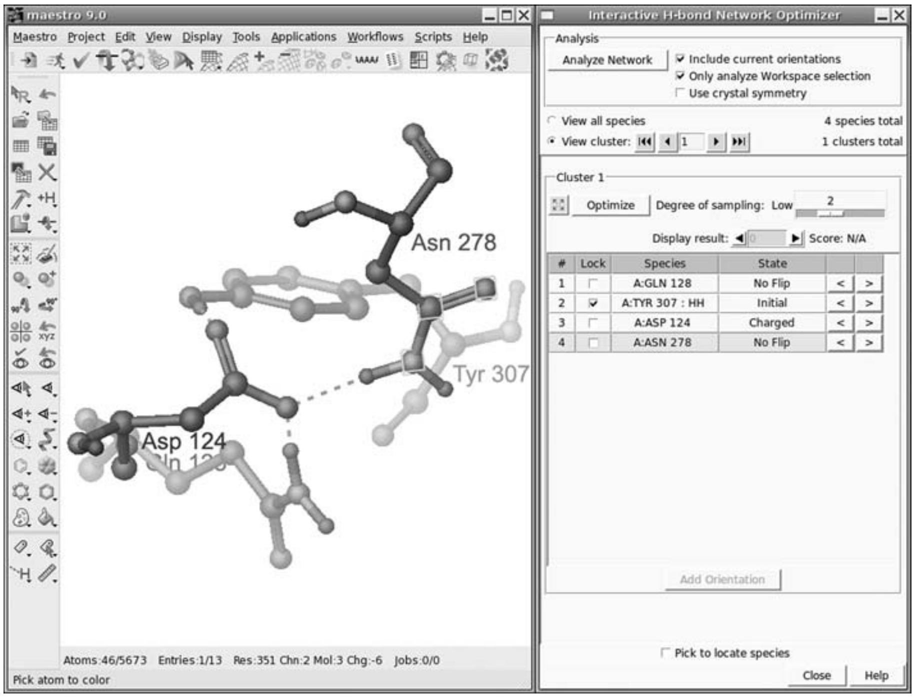
2D structure height: 200

Entries: 13 total, 13 shown, 0 selected, 0 included   Groups: 0 total, 0 selected

Figure 18.10.2.1
The project table entries track the steps of this refinement in progress, using the model from PDB (Protein Data Bank; Berman *et al.*, 2000) entry 1mdq to solve the structure of entry 1mpd. The most important statistics for monitoring the course of the refinement are always easily available.

18.10.3.2. *Molecular models consistent with X-ray data*

All current programs for refinement of biological crystal structures employ geometric restraints that either originate with or are consistent with a survey of small-molecule structures by Engh & Huber (1991). Other geometric restraints for group planarity and non-bonded interactions have been added to each refinement package independently from different sources. *PrimeX* employs the internally consistent molecular description embodied in the OPLS force field to achieve


maestro 9.0
Maestro Project Edit View Display Tools Applications Workflows Scripts Help
Asn 278
Tyr 307
Asp 124
Gln 128
Interactive H-bond Network Optimizer
Analysis
Analyze Network
Optimize
Add Orientation
Close
Help

Atoms:46/5673 Entries:1/13 Res:351 Chn:2 Mol:3 Chg:-6 Jobs:0/0

Pick atom to color

☑ Include current orientations
☑ Only analyze Workspace selection
☐ Use crystal symmetry

○ View all species — 4 species total
● View cluster: 1 — 1 clusters total

Cluster 1

Degree of sampling: Low 2

Display result: 0 Score: N/A

#	Lock	Species	State		
1	☐	A:GLN 128	No Flip	<	>
2	☑	A:TYR 307 : HH	Initial	<	>
3	☐	A:ASP 124	Charged	<	>
4	☐	A:ASN 278	No Flip	<	>

☐ Pick to locate species

Figure 18.10.3.1

The hydrogen-bond network optimizer clusters hydrogen-bond donors and acceptors to group together those that could possibly interact, so that these interactions can be efficiently optimized. Individual clusters may be examined and optimized by the user as shown, or all of the clusters may be optimized in a single operation.

the same goal of restraining stereochemistry to values typical of small-molecule structures.

During refinement, all crystallographic calculations in *PrimeX* are made without any contribution from hydrogen atoms. This strategy avoids exacerbation of the problem of overfitting that would occur if hydrogen coordinates were also considered to be adjustable parameters in refinement. Rather, the hydrogen-atom coordinates in the model can be thought of as a means of applying the more accurate molecular description found in the OPLS force field, while the positions of non-hydrogen atoms are influenced by the crystallographic gradients. This distinct hybrid approach for the calculation of the two types of terms during refinement works well, despite the seeming asymmetry of the treatment (J. A. Bell, unpublished work).

18.10.3.3. Automation of refinement tasks

As well as executing through the graphical interface, most *PrimeX* tasks can be run through command-line execution, so these tasks can be easily integrated into workflows. Future development will be focused on combining multiple tasks and diagnostic information into larger automated tasks that will reduce the amount of routine work required from users during refinement.

18.10.4. *PrimeX* implementation and theory

18.10.4.1. Force-field model and automatic atom typing

An important feature of *PrimeX* is that all molecules are automatically treated with atom typing and parameters are generated without the intervention of the crystallographer. This capability saves a large amount of time that might otherwise be spent collecting this information from disparate sources. The use of OPLS in *PrimeX* thus ensures that all ligand and protein parameters are similarly scaled and are consistent between all parts of the structure.

18.10.4.2. Reciprocal-space calculations

18.10.4.2.1. Restrained maximum-likelihood minimization

For reciprocal-space refinement with a maximum-likelihood target, the formulation of Pannu & Read (1996) was implemented following the general concepts developed in the publications of Brünger and co-workers (Brünger, 1989; Brünger *et al.*, 1998). A maximum-likelihood target has been shown to improve the convergence of refinement and to reduce the effects of model bias (Murshudov *et al.*, 1997). Optionally, a more traditional least-squares target may be employed. Three different minimizers are provided: conjugate gradient, quasi-Newton and truncated Newton (Zhu, Shirts, Friesner & Jacobson, 2007). Parts of the model related by noncrystallographic symmetry may be restrained to be similar with respect to both coordinates and displacement parameters during refinement. All or parts of the structure can be easily grouped for rigid-body coordinate refinement. Refinement will automatically account for alternate conformations. Alternate conformations may vary in size and may encompass from just one atom to many residues. Restrained isotropic *B*-factor refinement is also provided. *B* factors may optionally be grouped by main-chain and side-chain atoms for each residue. Grouped occupancy refinement is also available.

18.10.4.2.2. Simulated annealing

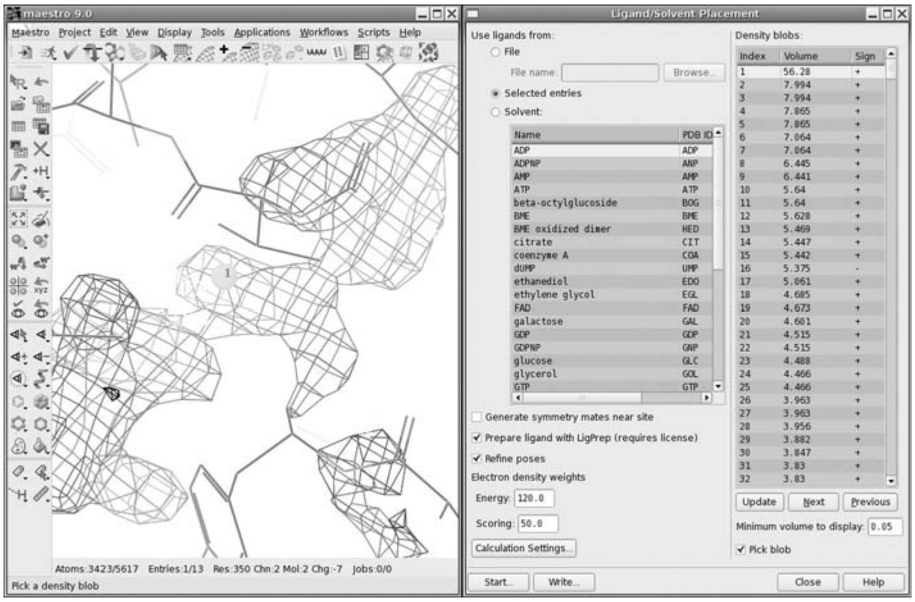
Simulated-annealing refinement was implemented within the general-purpose molecular modelling package *IMPACT* (Banks *et al.*, 2005), employing concepts for simulated-annealing refinement validated in the program *CNS* (Adams *et al.*, 1997). *PrimeX* simulated annealing provides two alternative energy models for dynamic simulation refinement. In the complete method, all molecular-mechanics terms are evaluated during the simulation. In the approximate method, the electrostatic and implicit solvation terms are not evaluated, a method similar to that employed in *CNS* (Adams *et al.*, 1997).

18.10.4.2.3. Map generation

Map calculations in *PrimeX* are based on the sigma*A* weighting scheme of Read (1986), a data treatment that has been shown to decrease the bias in electron-density maps. Optionally, an unweighted map may be calculated.

18.10.4.2.4. Omit maps

Omit maps (Bhat & Cohen, 1984; Bhat, 1988) have been implemented because of their demonstrated ability to reduce model-derived bias in maps (Vellieux & Dijkstra, 1997). The kicked omit maps described by Guncar *et al.* (2000) have also been included in *PrimeX*. After all the atoms to be omitted have been removed from the coordinates for the calculation, the model is modified by adding a small displacement of random size and direction to each coordinate before the phases are calculated. This process is repeated multiple times and the multiple maps are averaged. Either of these methods can be applied in a composite manner where small portions of the structure are omitted systematically over the entire map volume. Finally, simulated-annealing omit maps can be calculated by deleting the atoms to be omitted from the model, followed by simulated-annealing refinement and automatic map generation.



Use ligands from:

○ File

File name:

● Selected entries

○ Solvent:

Name	PDB ID
ADP	ADP
ADPNP	ANP
AMP	AMP
ATP	ATP
beta-octylglucoside	BOG
BME	BME
BME oxidized dimer	HED
citrate	CIT
coenzyme A	COA
dUMP	UMP
ethanediol	EDO
ethylene glycol	EGL
FAD	FAD
galactose	GAL
GDP	GDP
GDPNP	GNP
glucose	GLC
glycerol	GOL
GTP	GTP

☐ Generate symmetry mates near site

☑ Prepare ligand with LigPrep (requires license)

☑ Refine poses

Electron density weights

Energy: 120.0

Scoring: 50.0

Density blobs:

Index	Volume	Sign
1	56.28	+
2	7.994	+
3	7.994	+
4	7.865	+
5	7.865	+
6	7.064	+
7	7.064	+
8	6.445	+
9	6.441	+
10	5.64	+
11	5.64	+
12	5.628	+
13	5.469	+
14	5.447	+
15	5.442	+
16	5.375	-
17	5.061	+
18	4.685	+
19	4.673	+
20	4.601	+
21	4.515	+
22	4.515	+
23	4.488	+
24	4.466	+
25	4.466	+
26	3.963	+
27	3.963	+
28	3.956	+
29	3.882	+
30	3.847	+
31	3.83	+
32	3.83	+

Minimum volume to display: 0.05

☑ Pick blob

Atoms:3423/5617 Entries:1/13 Res:350 Chn:2 Mol:2 Chg:-7 Jobs:0/0

Pick a density blob

Figure 18.10.4.1

The ligand/solvent placement GUI provides a list of electron-density blobs from a map with coefficients $F_o - F_c$, together with their volume, and several ways to specify the molecules to be placed into the electron density. Multiple molecules may be included in a single placement operation at a single site. Ligand-placement options include the generation of charge and tautomer states of the ligand with the program *LigPrep* and the inclusion of symmetry-related molecules in the calculation at the ligand site.

18.10.4.3. *Real-space tools*

18.10.4.3.1. *Loop refinement*

Polypeptide segments, both internal (loops) and external (tails), can be built from the target sequence as stored in the program. The program builds polypeptides according to an algorithm derived from the protein-modelling program *Prime* (Jacobson *et al.*, 2004; Zhu, Shirts & Friesner, 2007). Using the end of the existing structure as an anchor point, a residue is added and sampled in various conformations. All conformations that demonstrate even a modest fit to the electron density are kept. During the next addition, all stored conformations are used as starting points for sampling the conformation of the next residue. The total number of conformations may reach many thousands as additional residues are added. When all amino acids have been added, the conformations are clustered to reduce the redundancy of the set, treated with real-space minimization, scored by conformation (OPLS energy) and fitted to the electron density. When building loops, the structures are built from both ends of the structure gap. Only those pairs of built polypeptides that meet to form a contiguous loop are clustered, refined and evaluated. Typically, one to three of the highest scoring structures are returned by the program, depending on the quality of the electron density.

18.10.4.3.2. *Side-chain placement*

Also based on *Prime* technology, one or multiple side chains may be simultaneously evaluated and placed into electron density. Even if two side chains have overlapping conformations at the start of the process, they will both be placed into appropriate electron density. As an alternative, various common rotamers of side chains may be explored interactively in *Maestro* and the best fit to the electron density selected visually or automatically.

18.10.4.3.3. *Minimization*

Real-space minimization provides an additional method of optimizing model coordinates according to the electron density and force field, as well as playing an integral part in the building of polypeptide chains, the placement of amino-acid side chains and the placement of ligands.

18.10.4.4. *Water placement*

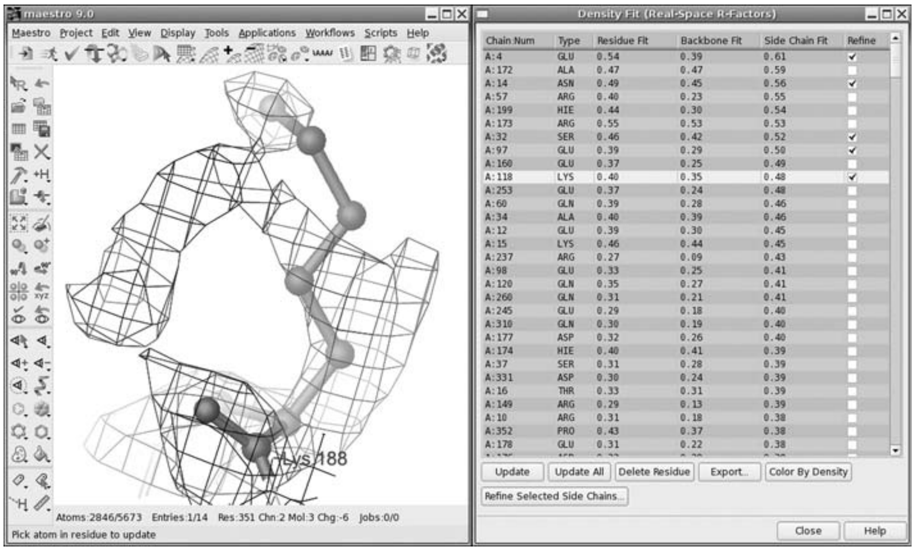
Waters may be added by *PrimeX* to difference electron-density maps for positive peaks exceeding a selected level of significance. Prospective water sites in the electron-density map are screened by three different criteria related to distance from other atoms. Water molecules that refine to a B factor above a selectable cutoff are deleted.

18.10.4.5. *Ligand placement*

Ligand placement can be accessed through a convenient GUI (Fig. 18.10.4.1). The algorithm for ligand placement was derived from the docking program *Glide* (Friesner *et al.*, 2004). In brief, the program abstracts the protein model as an energy grid in which the properties of the residues are encoded. Multiple conformations are generated for the ligand, and then a search is made for the best possible position and orientation in the vicinity of the electron density that the user has identified. The many alternative solutions are reduced in number with the application of a dual scoring system: (i) the GlideScore is an evaluation of the chemical complementarity between the ligand and the protein; (ii) the DensScore is a measure of the fit of the ligand to the observed electron density. Both of these scores are used to rank the fitting of each ligand conformation in each orientation. The weighting of the DensScore *versus* the GlideScore is under user control and it strongly favours the DensScore by default. Surviving candidates are subjected to rigid-body and torsion-angle refinement against an energy function that includes the OPLS energy and the DensScore. The weighting of the DensScore *versus* the OPLS energy is also under user control, and it strongly favours the DensScore by default. Finally, the top few dozen candidates are treated with real-space refinement and are ranked by real-space R factor. The top five poses for the ligand are returned with the protein structure for evaluation by the user.

18.10.4.6. *Validation*

PrimeX provides three main validation tools. 'Protein reports' lists geometric variations from expected values in table form. This table can be sorted by residue or by the magnitude of the deviation, to highlight parts of the structure most in need of attention. A density fit table (Fig. 18.10.4.2) lists the real-space R factors for each main chain and side chain and for the entire residue. This table can also be sorted by any column. Clicking on any row of either table centres the molecular display on that


maestro 9.0
Maestro Project Edit View Display Tools Applications Workflows Scripts Help
Lys188
Atoms:2846/5673 Entries:1/14 Res:351 Chn:2 Mol:3 Chg:-6 Jobs:0/0
Pick atom in residue to update
Density Fit (Real-Space R-Factors)
Update
Update All
Delete Residue
Export...
Color By Density
Refine Selected Side Chains...
Close
Help

Chain:Num	Type	Residue Fit	Backbone Fit	Side Chain Fit	Refine
A:4	GLU	0.54	0.39	0.61	✓
A:172	ALA	0.47	0.47	0.59	☐
A:14	ASN	0.49	0.45	0.56	✓
A:57	ARG	0.40	0.23	0.55	☐
A:199	HIE	0.44	0.30	0.54	☐
A:173	ARG	0.55	0.53	0.53	☐
A:32	SER	0.46	0.42	0.52	✓
A:97	GLU	0.39	0.29	0.50	✓
A:160	GLU	0.37	0.25	0.49	☐
A:118	LYS	0.40	0.35	0.48	✓
A:253	GLU	0.37	0.24	0.48	☐
A:60	GLN	0.39	0.28	0.46	☐
A:34	ALA	0.40	0.39	0.46	☐
A:12	GLU	0.39	0.30	0.45	☐
A:15	LYS	0.46	0.44	0.45	☐
A:237	ARG	0.27	0.09	0.43	☐
A:98	GLU	0.33	0.25	0.41	☐
A:120	GLN	0.35	0.27	0.41	☐
A:260	GLN	0.31	0.21	0.41	☐
A:245	GLU	0.29	0.18	0.40	☐
A:310	GLN	0.30	0.19	0.40	☐
A:177	ASP	0.32	0.26	0.40	☐
A:174	HIE	0.40	0.41	0.39	☐
A:37	SER	0.31	0.28	0.39	☐
A:331	ASP	0.30	0.24	0.39	☐
A:16	THR	0.33	0.31	0.39	☐
A:149	ARG	0.29	0.13	0.39	☐
A:10	ARG	0.31	0.18	0.38	☐
A:352	PRO	0.43	0.37	0.38	☐
A:178	GLU	0.31	0.22	0.38	☐

Figure 18.10.4.2
The density fit table provides a convenient method for finding residues where the main chain or side chain in the model poorly fits the electron density. Clicking on the line for Lys118 in chain A centres that residue in the workspace display, allowing the user to bring to bear the appropriate tool to resolve the problem.

residue. An interactive Ramachandan plot provides additional analysis of main-chain structure.

18.10.5. Conclusion

PrimeX is a combination of well established methodology with innovations focused on producing the most useful model from the refinement of biological molecules. Future directions to be explored will include TLS scaling (Winn *et al.*, 2003), increasing levels of automation, more reliable water placement, and refinement of ensembles of structures as a means of representing the flexibility of the molecule and the uncertainty of the atomic positions (Knight *et al.*, 2008).

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Banks, J. L., Beard, H. S., Cao, Y., Cho, A. E., Damm, W., Farid, R., Felts, A. K., Halgren, T. A., Mainz, D. T., Maple, J. R., Murphy, R., Philipp, D. M., Repasky, M. P., Zhang, L. Y., Berne, B. J., Friesner, R. A., Gallicchio, E. & Levy, R. M. (2005). *Integrated Modeling Program, Applied Chemical Theory (IMPACT). J. Comput. Chem.* **26**, 1752–1780.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Bhat, T. N. (1988). *Calculation of an OMIT map. J. Appl. Cryst.* **21**, 279–281.

Bhat, T. N. & Cohen, G. H. (1984). *OMITMAP: an electron density map suitable for the examination of errors in a macromolecular model. J. Appl. Cryst.* **17**, 244–248.

Brünger, A. T. (1989). *A memory-efficient fast Fourier transformation algorithm for crystallographic refinement on supercomputers. Acta Cryst.* A**45**, 42–50.

Brünger, A. T. (1992). *X-PLOR.* Version 3.1. *A system for X-ray crystallography and NMR*. Yale University, Connecticut, USA.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Cryst.* D**54**, 905–921.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P. & Shenkin, P. S. (2004). *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J. Med. Chem.* **47**, 1739–1749.

Guncar, G., Klemencic, I., Turk, B., Turk, V., Karaoglanovic-Carmona, A., Juliano, L. & Turk, D. (2000). *Crystal structure of cathepsin X: a flip-flop of the ring of His23 allows carboxy-monopeptidase and carboxy-dipeptidase activity of the protease. Structure*, **8**, 305–313.

Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E. & Friesner, R. A. (2004). *A hierarchical approach to all-atom protein loop prediction. Proteins*, **55**, 351–367.

Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids*. *J. Am. Chem. Soc.* **118**, 11225–11236.

Kaminski, G., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). *Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides*. *J. Phys. Chem. B*, **105**, 6474–6487.

Knight, J. L., Zhou, Z., Gallicchio, E., Himmel, D. M., Friesner, R. A., Arnold, E. & Levy, R. M. (2008). *Exploring structural variability in X-ray crystallographic models using protein local optimization by torsion-angle sampling. Acta Cryst.* D**64**, 383–396.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Vellieux, F. M. D. & Dijkstra, B. W. (1997). *Computation of Bhat's OMIT maps with different coefficients. J. Appl. Cryst.* **30**, 396–399.

Winn, M. D., Murshudov, G. N. & Papiz, M. Z. (2003). *Macromolecular TLS refinement in REFMAC at moderate resolutions. Methods Enzymol.* **374**, 300–321.

Zhu, K., Shirts, M. R. & Friesner, R. A. (2007). *Improved methods for side chain and loop predictions via the protein local optimization program: variable dielectric model for implicitly improving the treatment of polarization effects*. *J. Chem. Theory Comput.* **3**, 2108–2119.

Zhu, K., Shirts, M. R., Friesner, R. A. & Jacobson, M. P. (2007). *Multiscale optimization of a truncated Newton minimization algorithm and application to proteins and protein–ligand complexes*. *J. Chem. Theory Comput.* **3**, 640–648.

International Tables for Crystallography (2012). Vol. F, Chapter 18.11, pp. 539–547.

Chapter 18.11. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution

P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart

18.11.1. Foundations

18.11.1.1. PHENIX architecture

The *PHENIX* (Adams *et al.*, 2002) architecture is designed from the ground up as a hybrid system of tightly integrated interpreted ('scripted') and compiled software modules. A mix of scripted and compiled components is invariably found in all major successful crystallographic packages, but often the scripting is added as an afterthought in an *ad hoc* fashion using tools that predate the object-oriented programming era. While such *ad hoc* systems are quickly established, they tend to become a severe maintenance burden as they grow. In addition, users are often forced into many time-consuming routine tasks such as manually converting file formats. In *PHENIX*, the scripting layer is the heart of the system. With only a few exceptions, all major functionality is implemented as modules that are exclusively accessed *via* the scripting interfaces. The object-oriented Python scripting language (Lutz & Ascher, 1999) is used for this purpose. In about two decades, a large developer/user community has produced millions of lines of highly uniform, interoperable, mature and openly available sources covering all aspects of programming ranging from simple file handling to highly sophisticated network communication and fully featured cross-platform graphical interfaces. Embedding crystallographic methods into this environment enables an unprecedented degree of automation, stability and portability. By design, the object-oriented programming model fosters shared collaborative development by multiple groups. It is routine practice to hierarchically recombine modules written by different groups into ever more complex procedures that appear uniform from the outside. A more detailed overview of the key software technology leading to all these advances, presented in the context of crystallography, can be found in Grosse-Kunstleve *et al.* (2002).

In addition to the advantages outlined in the previous paragraph, the scripting language is generally most efficient for the rapid development of new algorithms. However, runtime performance considerations often dictate that numerically intensive calculations are eventually implemented in a compiled language. The first choice of a compiled language is of course to reuse the same language environment as used for the scripting language itself, which is a C/C++ environment. Not only is this the mainstream software environment on all major platforms used today, but with probably hundreds of millions of lines of C/C++ sources in existence it is an environment that is virtually guaranteed to thrive in the long term. An in-depth discussion of the combined use of Python and C++ can be found in Grosse-Kunstleve *et al.* (2002) and Abrahams & Grosse-Kunstleve (2003). This model is used throughout the *PHENIX* system.

18.11.1.2. Graphical user interface

A new graphical user interface (GUI) for *PHENIX* was introduced in version 1.4. It uses the open-source wxPython toolkit, which provides a 'native' look on each operating system. Development has focused on providing interfaces around the existing command-line programs with minimal modification, using the same underlying configuration system (*libtbx.phil*; Grosse-Kunstleve *et al.*, 2005) as used by most *PHENIX* programs as a template to automatically generate controls. Because these programs are implemented primarily as Python modules, complex data including models, reflections and other viewable data may be exchanged with the GUI without resorting to parsing log files. The current *PHENIX* release (version 1.5) includes GUIs for *phenix.refine* (Afonine *et al.*, 2005), *phenix.xtriage* (Zwart *et al.*, 2005), the *AutoSol* (Terwilliger *et al.*, 2009), *AutoBuild* (Terwilliger, Grosse-Kunstleve, Afonine, Moriarty, Adams *et al.*, 2008) and *LigandFit* (Terwilliger *et al.*, 2006) wizards, the restraints editor *REEL*, all of the validation tools and several utilities for creating and manipulating maps and reflection files. More recent builds of *PHENIX* contain a new GUI for the *AutoMR* wizard and future releases will include a new interface for *Phaser* (McCoy *et al.*, 2007).

Intrinsically graphical data is visualized with embedded graphs (using the free matplotlib Python library) or a simple OpenGL viewer. This simplifies the most complex parameters, such as atom selections in *phenix.refine*, which can be visualized or picked interactively with the built-in viewer. The GUI also serves as a platform for additional automation and user customization. Similarly to the *CCP*4 interface (*CCP*4*i*; Potterton *et al.*, 2003), *PHENIX* manages data and task history for separate user-defined projects. Default parameters and input files can be specified for each project; for instance, the generation of ligand restraints from the *phenix.refine* GUI gives the user the option of automatically loading these restraints in future runs.

The popularity of Python as a scientific programming language has led to its use in many other structural-biology applications, especially molecular-graphics software. The *PHENIX* GUI includes extension modules for the modeling programs *Coot* (Emsley & Cowtan, 2004; Chapter 17.1) and *PyMOL* (DeLano, 2002), both of which are controlled remotely from *PHENIX* using the XML-RPC protocol. This allows the interfaces to integrate seamlessly; any model or map in *PHENIX* can be automatically opened in *Coot* with a single click. In programs that iteratively rebuild or refine structures, such as *AutoBuild* and *phenix.refine*, the current model and maps will be continually updated in *Coot* and/or *PyMOL* as soon as they are available. In the validation utilities, clicking on any atom or residue flagged for poor statistics will recentre the graphics windows on that atom. Remote control of the *PHENIX* GUI is also simple, using the same protocol, and simple extensions to the *Coot* interface provide direct launching of *phenix.refine* with a model preloaded.

Copyright © 2012 International Union of Crystallography

18.11.2. Analysis of experimental data

PHENIX has a range of tools for the analysis, validation and manipulation of X-ray diffraction data. A comprehensive tool for analysing X-ray diffraction data is *phenix.xtriage* (Zwart *et al.*, 2005), which carries out tests ranging from space-group determination and detection of twinning to detection of anomalous signal. These tests provide the user and the various wizards with a set of statistics that characterize a data set. For analysis of twinning, *phenix.xtriage* consolidates a number of statistics to provide a balanced verdict of possible symmetry and twin-related issues with the data. *Phenix.xtriage* provides the user with feedback on the overall characteristics of the data. Routine usage of *phenix.xtriage* during or immediately after data collection has resulted in the timely discovery of twinning or other issues (Flynn *et al.*, 2007; Kostelecky *et al.*, 2009). Detection of these idiosyncrasies in the data typically reduces the overall effort in a successful structure determination.

A likelihood-based estimation of the overall anisotropic scale factor is performed using the likelihood formalism described by Popov & Bourenkov (2003). Database-derived standard Wilson plots for proteins and nucleic acids are used to detect anomalies in the mean intensity. These anomalies may arise from ice rings or other issues (Morris *et al.*, 2004). Data strength and low-resolution completeness are also analysed. The presence of anomalous signal is detected by analysis of the measurability, a quantity expressing the fraction of statistically significant Bijvoet differences in a data set (Zwart, 2005). The native Patterson function is used to detect the presence of pseudo-translational symmetry. A database-derived empirical distribution of maximum peak heights is used to assign significance to detected peaks in the Patterson function.

A comprehensive automated twinning analysis is performed. Twin laws are derived from first principles to facilitate the identification of pseudo-merohedral cases. Amplitude and intensity ratios, $\langle |E^2 - 1| \rangle$ values, the L statistic (Padilla & Yeates, 2003) and $N(Z)$ plots are derived from data cut to the resolution limit suggested by the data-strength analysis. The removal of shells of data with relatively high noise content greatly improves the automated interpretation of these statistics. A Britton plot, H test and a likelihood-derived approach are used to estimate twin fractions when twin laws are present. If a model has been supplied, an R *versus* R (Lebedev *et al.*, 2006) analysis is carried out. This type of analysis is of particular use when dealing with pseudo-symmetry, space-group problems and twinning (Zwart *et al.*, 2008).

To test for inconsistent indexing between different data sets, a set of reindexing laws is derived from first principles given the unit cells and space groups of the sample and reference data sets. A correlation analysis suggests the most likely choice of re-indexing of the data. Analysis of the metric symmetry of the unit cell provides a number of likely point groups. A likelihood-inspired method is used to suggest the most likely point group of the data. Subsequent analysis of systematic absences in a likelihood framework ranks subsequent space-group possibilities (details to be published).

18.11.3. Substructure determination, phasing and molecular replacement

After ensuring that the diffraction data are sound and understood, the next critical necessity for solving a structure is the determination of phases using one of several strategies (Adams, Afonine *et al.*, 2009).

18.11.3.1. Substructure determination

The substructure-determination procedure implemented as *phenix.hyss* (*Hybrid Substructure Search*; Grosse-Kunstleve & Adams, 2003) combines the multi-trial dual-space recycling approaches pioneered by *Shake-and-Bake* (Miller *et al.*, 1994) and later *SHELXD* (Sheldrick, 2008) with the use of the fast translation function (Navaza & Vernoslova, 1995; Grosse-Kunstleve & Brunger, 1999). The fast translation function is the basis for a systematic search in the Patterson function (performed in reciprocal space), in contrast to the stochastic alternative of *SHELXD* (performed in direct space). *Phenix.hyss* is the only substructure-determination program to fully integrate automatic comparison of the substructures found in multiple trials *via* a Euclidean model matching procedure (part of the *cctbx* open-source libraries). This allows *phenix.hyss* to detect whether the same solution was found multiple times and to terminate automatically if this is the case. Extensive tests with a variety of single-wavelength anomalous diffraction (SAD) data sets (Grosse-Kunstleve & Adams, 2003) have led to a parameterization of the procedure that balances runtime considerations and the likelihood that repeated solutions present the correct substructure. In many cases the procedure finishes in seconds if the substructure is detectable from the input data.

18.11.3.2. Phasing

Phaser, available in *PHENIX* as *phenix.phaser*, applies the principle of maximum likelihood to solving crystal structures by molecular replacement, by SAD or by a combination of both. The likelihood targets take proper account of the effects of different sources of error (and, in the case of SAD phasing, their correlations) and allow different sources of information to be combined. In solving a molecular-replacement problem with a number of different components, the information gained from a partial solution increases the signal in the search for subsequent components. Because the likelihood scores for different models can be directly compared, decisions among models can readily be made as part of automation strategies (discussed below).

18.11.3.3. Noncrystallographic symmetry (NCS)

Noncrystallographic symmetry is an important feature of many macromolecular crystals that can be used to greatly improve electron-density maps. *PHENIX* has tools for the identification of NCS and for using NCS and multiple crystal forms of a macromolecule in phase improvement.

Phenix.find_ncs and *phenix.simple_ncs_from_pdb* are tools for the identification of noncrystallographic symmetry in a structure using information from a heavy-atom substructure or an atomic model. *Phenix.simple_ncs_from_pdb* will identify NCS and generate transformations from the chains in a model in a PDB file. *Phenix.find_ncs* will identify NCS from either a heavy-atom substructure (Terwilliger, 2002*a*) or the chains in a PDB file and will then compare this NCS with the density in a map to verify that the NCS is actually present.

Phenix.multi_crystal_average is a method for combining information from several crystal forms of a structure. It is especially well suited to cases where each crystal form has its own NCS, adjusting phases for each crystal form so that all the NCS copies in all crystals are as similar as possible.

NCS restraints should normally be applied in density modification and model building in all cases except where there is clear evidence that NCS is not present. In density modification within *PHENIX* the presence of NCS is identified from the heavy-atom sites or from an atomic model if available. The local correlation of density in NCS-related locations is then used automatically to set variable restraints on NCS symmetry in the map. In refinement, NCS symmetry is applied through coordinate restraints, targeting the positions of each NCS copy relative to those of the other NCS-related chains. The default NCS restraints in *PHENIX* are very tight, with targets of 0.05 Å r.m.s. At resolutions lower than about 2.5 Å these tight restraints on NCS should usually be applied. At higher resolutions it may be appropriate to use looser restraints or to remove them altogether. Additionally, if there are segments of the chains that clearly do not obey the NCS relationships they should be excluded from the NCS restraints. Normally this is performed automatically, but it can also be specified explicitly.

18.11.4. Model building, ligand fitting and nucleic acids

Key steps in the analysis of a macromolecular crystal structure are building an initial core model, identification and fitting of ligands into the electron-density map and building an atomic model for loop regions that are less well defined than the majority of the structure. *PHENIX* has tools for rapid model building of secondary structure and main-chain tracing (*phenix.find_helices_strands*) and for the fitting of flexible ligands (*phenix.ligandfit*) as well as for fitting a set of ligands to a map (*phenix.find_all_ligands*) and for the identification of ligands in a map (*phenix.ligand_identification*). *PHENIX* additionally has a tool for the fitting of missing loops (*phenix.fit_loops*). Validation tools are provided so that the models produced can be validated at each step along the way.

18.11.4.1. Model building

Phenix.find_helices_strands will rapidly build a secondary-structure-only model into a map or very rapidly trace the polypeptide backbone of a model into a map. To build secondary structure in a map, *phenix.find_helices_strands* identifies α-helical regions and β-strand segments, models idealized helices and strands into the corresponding density, allowing for bending of the helices and strands, and assembles these into a composite model. To very rapidly trace the main chain in a map, *phenix.find_helices_strands* finds points along ridgelines of high density where Cα atoms might be located, identifies pairs and then triplets of these Cα atoms that have density between the atoms and plausible geometry, constructs all possible connections of these Cα atoms into nonamers and then identifies all the longest possible chains that can be made by joining the nonamers. This process can build a Cα model at a rate of about 20 residues per second, yielding a backbone model that can readily be interpreted visually or automatically to evaluate the quality of the map that it is based on.

Phenix.fit_loops will fit missing loops in an atomic model. It uses *RESOLVE* model building (Terwilliger, 2003*a*,*b*,*c*) to extend the chain from either end where a loop is missing and to connect the chains into a loop with the expected number of residues.

18.11.4.2. Ligand fitting

Phenix.ligandfit is a tool for fitting a flexible ligand into an electron-density map (Terwilliger *et al.*, 2006). The key approaches used are breaking the ligand into its component rigid-body parts, finding where each of these can be placed into density, tracing the remainder of the ligand based on the positions of these core rigid-body parts and re-combining the best parts of multiple fits while scoring based on the fit to the density.

Phenix.find_all_ligands is a tool for finding all the instances of each of several ligands in an electron-density map. *Phenix.find_all_ligands* finds the largest contiguous region of unused density in a map and uses *phenix.ligandfit* to fit each supplied ligand into that density. It then chooses the ligand that has the highest real-space correlation to the density (Terwilliger, Adams *et al.*, 2007). It then repeats this process until no ligands can be satisfactorily fitted into any remaining density in the map.

Phenix.ligand_identification is a tool for identifying which ligands are compatible with unknown electron density in a map (Terwilliger, Adams *et al.*, 2007). It can search using the 200 most common ligands from the PDB or from a user-supplied list of ligands. *Phenix.ligand_identification* uses *phenix.ligandfit* to fit each ligand to the map and identifies the best-fitting ligand using the real-space correlation and surface complementarity of the ligand and the atoms in the structure surrounding the ligand-binding site.

18.11.4.3. RNA and DNA

In common with most macromolecular crystallographic tools, *PHENIX* was originally developed with protein structures primarily in mind. Now that nucleic acids, and especially RNA, are increasingly important in large biological structures, the system is being modified in places where subtle differences in procedure are needed rather than just the relevant libraries.

Model building in *phenix.autobuild* now has a preliminary set of nucleic acid procedures that take advantage of the relatively well determined phosphate and base positions, as well as the preponderance of double helix, and in the future will make use of the RNA backbone conformers recently defined by the RNA Ontology Consortium (Richardson *et al.*, 2008). Nucleic acid structures benefit significantly from torsion-angle refinement, which has recently been added to the options in *phenix.refine*. A principal problem in RNA models is getting the ribose pucker correct, although it is known to consist almost entirely of either C3′-*endo* (which is commoner and that found in the A-form helix) or C2′-*endo* (Altona & Sundaralingam, 1972). *MolProbity* uses the perpendicular distance from the 3′ phosphate to the line of the C1′—N1/9 glycosidic bond as a reliable diagnostic of ribose pucker (Davis *et al.*, 2007; Chapter 21.6). This same test has now been built into *phenix.refine* to allow the use of pucker-specific target parameters for bond lengths, angles and torsions (Gelbin *et al.*, 1996) rather than the uneasy compromise values (Parkinson *et al.*, 1996) used in most pucker-agnostic refinement. Currently, if an incorrect pucker is diagnosed it must usually be fixed by user rebuilding, for instance in *Coot* (Emsley & Cowtan, 2004; Chapter 17.1) or in *RNABC* (Wang *et al.*, 2008). A rebuilding functionality will probably be incorporated into *PHENIX* soon, but in the meantime the refinement will now correctly maintain the geometry of a C2′-*endo* pucker once it has been built and identified using conformation-specific residue names.

18.11.4.4. Maps, models and avoiding bias

Phenix.refine (and the graphical tool *phenix.maps*) can produce various types of maps, including anomalous difference, maximum-likelihood weighted ($p{*}mF_{\rm obs} - q{*}DF_{\rm model})\exp(i\alpha_{\rm model})$ and regular $(p{*}F_{\rm obs} - q{*}F_{\rm model}) \times \exp(i\alpha_{\rm model})$, where p and q are any user-defined numbers, filled and kick maps. The coefficients m and D of likelihood-weighted maps (Read, 1986) are computed using test-set reflections as described in Lunin & Skovoroda (1995) and Urzhumtsev *et al.* (1996).

Data incompleteness, especially systematic incompleteness, can cause map distortions (Lunin, 1988; Tronrud, 1997). An approach to remedying this problem is to replace ('fill') missing observations with nonzero values. One can use $DF_{\rm model}$ (similarly to *REFMAC*; Murshudov *et al.*, 1997) to replace the missing $F_{\rm obs}$ or use $\langle F_{\rm obs}\rangle$, where the $F_{\rm obs}$ are averaged across a resolution bin around the missing $F_{\rm obs}$ value. Based on a limited number of tests, both 'filling' schemes produce similar results, reiterating the importance of phases. However, it is important to keep in mind that by replacing missing $F_{\rm obs}$ there is a risk of introducing bias and, obviously, the more incomplete the data are the larger the risk. At present it is advisable to use both maps simultaneously: filled and not filled.

An average kick map (AK map; Gunčar *et al.*, 2000; Turk, 2007; Pražnikar *et al.*, 2009) is the result of the following procedure. A large ensemble of structures is created where the coordinates of each structure from the ensemble are all randomly shaken. A map is then computed for each structure. Finally, all maps are averaged to generate one AK map. An AK map is expected to have less bias and less noise and to enhance the existing signal, and can potentially clarify some initially bad densities.

A computationally intensive but powerful method of creating a very low bias map is to carry out iterative model building and refinement while omitting one region of the map from all calculations of structure factors (Terwilliger, Grosse-Kunstleve, Afonine, Moriarty, Adams *et al.*, 2008). The *phenix.autobuild* iterative-build OMIT map procedure carries this out automatically for either a single OMIT region or for overlapping OMIT regions to create a composite iterative-build OMIT map.

18.11.5. Model, and model-to-data, validation

The result of crystallographic structure determination is the atomic model. There are three principal components in assessing model quality: the covalent model geometry, the model stereochemistry and the quality of fit between the model and experimental data in both real space and in reciprocal space. All three provide overall measures, and the first two plus the real-space aspect of the third also provide checks for local outliers, which give the best leverage for user intervention to actively improve model accuracy (Arendall *et al.*, 2005). (Validation of the experimental data was described in Section 18.11.2 above.) *PHENIX* includes many individual tools for specific aspects of validation, plus several systems that combine those results into overall summaries. Validation is provided both for user evaluation of the progress and results of a structure solution and also to help inform the automated choices made by other parts of the system.

Most aspects of the *MolProbity* model-validation tools (Davis *et al.*, 2007; Chapter 21.6) have been adapted or rewritten for integrated use within *PHENIX* and are presented to the user by the new GUI (Section 18.11.1.2). H atoms are added by *phenix.reduce*, with optimization of entire local hydrogen-bond networks, consideration of the first layer of crystallographic waters and optional correction of side-chain amide or histidine 180° 'flips' (Word, Lovell, Richardson & Richardson, 1999). All-atom contacts (Word, Lovell, LaBean *et al.*, 1999) are calculated by *phenix.probe*, which provides the atomic overlap information needed for the validation of serious all-atom steric clashes, and can also be visualized in *Coot*. For the *PHENIX* GUI, the set of *MolProbity*-based tools provides both overall model statistics, such as clashscore and percentage of outliers, and detailed lists of the Ramachandran (Lovell *et al.*, 2003), rotamer (Lovell *et al.*, 2000), Cβ deviation (Lovell *et al.*, 2003) and clash outliers. The GUI also provides access to comprehensive validation kinemage graphics visualized in three dimensions in the graphical package *KiNG* (Chen *et al.*, 2009; Chapter 21.5), now included in *PHENIX*. Command-line tools are available for these validation methods: *phenix.rotalyze*, *phenix.ramalyze*, *phenix.cbetadev*, *phenix.clashscore*, *phenix.rna_validate*, *phenix.kinemage* (multi-criterion kinemage generation), *phenix.king*, *phenix.reduce* and *phenix.probe*. Additionally, *phenix.validate_model*, which analyses the deviations of bond lengths, bond angles, planarity *etc.* from ideal library values, complements the *MolProbity* torsional and atomic clash tools.

Phenix.real_space_correlation asserts the local model-to-data correspondence by providing a quantitative measure of how the atomic model fits the electron-density map at the residue or atom level (depending on the resolution). Rapidly obtaining a snapshot of global figures of merit for a crystallographic model and associated experimental data is a frequent task that is performed at all stages of structure solution. This task can be complicated for several reasons: the presence of novel ligands or nonstandard residues in the PDB-format (Berman *et al.*, 2000) coordinate file, data collected from twinned crystals, various reflection data file formats, different representation of atomic displacement parameters in the presence of TLS (Schomaker & Trueblood, 1968), experimental data type (X-ray and/or neutron), and files with multiple models and various formatting issues. *Phenix.model_vs_data* (Afonine, Grosse-Kunstleve *et al.*, 2010) is designed to automatically handle all these complications with minimal user input (a PDB file and a reflection data file) and provide a concise summary output. *Phenix.polygon* (Urzhumtseva *et al.*, 2009) is a graphical tool that is designed to indicate the similarity of validation parameters, such as free R value, for a particular structure compared with those deposited in the PDB. This comparison is performed for all other structures solved at similar resolution limits. The result is presented graphically. *Phenix.validation* combines all of the tools described above in one GUI, providing a single place for assessing the results of structure determination.

18.11.5.1. Model and structure-factor manipulation and analysis

PHENIX has a range of tools for displaying, analysing and manipulating structure-factor and model information. *Phenix.mtz.dump* and *phenix.cif_as_mtz* display and convert structure-factor data. *Phenix.print_sequence*, *phenix.pdb_atom_selection* and *phenix.pdbtools* display and manipulate coordinate files.

Phenix.tls is a tool for the extraction and manipulation of TLS information. Using this tool, TLS matrices and selections can be extracted from *REFMAC*- or *PHENIX*-formatted PDB file headers and the total or residual atomic B factors can be computed and output. Future functionality will include the complete analysis of TLS matrices and their graphical visualization.

Phenix.get_cc_mtz_mtz and *phenix.get_cc_mtz_pdb* are tools for analysing the agreement between maps based on a pair of MTZ files or between maps calculated from an MTZ file and a PDB file. The key attributes of these tools are that they automatically search all allowed origin shifts that might relate the two maps and that they write out a modified version of one of the MTZ files or of the PDB file, shifted to match the other.

18.11.6. Structure refinement

Phenix.refine is the state-of-the-art crystallographic structure-refinement engine of *PHENIX*. The foundational refinement machinery is a combination of highly efficient programming tools and new or rethought crystallographic algorithms. *Phenix.refine* possesses an extensive set of tools that cover the majority of refinement scenarios at any data resolution from low to ultra-high. Various reflection-data formats (for example, *CNS*, MTZ and *SHELX*) are recognized automatically. The input experimental data are checked for outliers (Read, 1999; Zwart *et al.*, 2005) and any reflections identified as such are excluded from the refinement calculations. Twinning can also be taken into account by providing a twin-law operator, which can be obtained using *phenix.xtriage*. Both X-ray and/or neutron diffraction data can be used and an option for joint XN refinement is available (simultaneous refinement against X-ray and neutron data; Adams, Mustyakimov *et al.*, 2009). Each refinement run begins with robust mask-based bulk-solvent correction and anisotropic scaling (Afonine *et al.*, 2005). Tools such as efficient rigid-body refinement (multiple-zones algorithm; Afonine *et al.*, 2009), simulated-annealing refinement (Brünger *et al.*, 1987) in Cartesian or torsion-angle space (Grosse-Kunstleve *et al.*, 2009), and automatic NCS detection and its use as restraints in refinement are important at low resolution and in the initial stages of refinement. A broad range of atomic displacement parameterizations are available, including grouped isotropic, constrained anisotropic (TLS) and individual atomic isotropic or anisotropic, allowing efficient modelling of atomic displacement parameters at any resolution (Afonine, Urzhumtsev *et al.*, 2010). Occupancy refinement (grouped, individual, group constrained for alternative conformations or any mixture) can be performed for any user-defined atoms. Atoms in alternative conformations are recognized automatically based on altLoc identifiers in the input PDB file and their occupancies are refined by default. Ordered solvent (water) model updating is integrated into the refinement process.

The availability of ultra-high-resolution data makes it possible to visualize the residual density arising from bonding effects; *phenix.refine* employs a novel interatomic scatterers model (Afonine *et al.*, 2007) to adequately account for these features. A flexible parameterization of H atoms allows their use at any resolution from subatomic (where their parameters can be refined individually) to low resolution (where a riding model is used). Refinement can be performed using a variety of refinement target functions, including maximum likelihood, maximum likelihood with experimental phase information and amplitude least squares. The refinement of coordinates can be performed in real or reciprocal space (allowing dual-space refinement). Novel ligands can easily be included in refinement by providing a corresponding CIF file as input (the CIF file can be automatically created using *phenix.ready_set*).

Manual fixing of amino-acid side-chain rotamers can be time-consuming, especially for large structures. Although the use of simulated-annealing refinement increases the convergence radius, it can still fail to fit incorrectly modelled side chains into the correct density. *Phenix.refine* has an option for automatic selection of the best rotamer based on a rotamer library (Lovell *et al.*, 2000) and optimal fit into the density (details to be published elsewhere). Furthermore, coupling real-space refinement with the built-in rotamer library and available *MolProbity* tools allows the automated identification and robust correction of common systematic errors involving backward-fit conformations for Leu, Thr, Val, Ile and Arg side chains, as developed and tested in the *Autofix* method (Headd *et al.*, 2009).

Phenix.refine allows multi-step complex refinement protocols in which most of the available refinement strategies can be combined with each other and applied to any selected part of the model. For example, a run of *phenix.refine* may perform rigid-body refinement, simulated annealing, individual and grouped B factors combined with TLS refinement, constrained occupancy refinement and automatic water picking.

The output of *phenix.refine* includes various maps (maximum-likelihood weighted, kicked, incompleteness corrected, anomalous difference and those with any user-defined coefficients), complete model and data statistics and a PDB file with a formatted REMARK 3 header ready for PDB deposition. The *phenix.refine* GUI is integrated with *Coot* and *PyMOL*, allowing seamless visual analysis of the refined model and associated maps.

Phenix.refine is tightly integrated with other *PHENIX* components, making structure solution, building and refinement a one-step process (for example, in the *AutoMR* and *AutoBuild* wizards). It is routinely tested by automatic re-refinement of all models in the PDB for which the experimental data are available.

18.11.6.1. Ligand-coordinate and restraint-geometry generation

The *electronic Ligand Builder and Optimization Builder* (*eLBOW*; Moriarty *et al.*, 2009) is a suite of tools designed for the reliable generation of Cartesian coordinates and geometry restraints for both novel and known ligands. In line with the rest of the *PHENIX* package, the *eLBOW* modules are written in Python, with the numerically intensive portions of the code written in C++. *eLBOW* is a flexible platform for converting a majority of common chemical inputs to optimized three-dimensional coordinates and geometry restraints for refinement. Ligand geometries can be minimized using the semi-empirical AM1 quantum-chemical method (Stewart, 2004), a numerically efficient and chemically accurate technique for the class of molecules commonly complexed with or bound to proteins.

In addition, a graphical user interface for editing geometry restraints and simple geometry manipulation of ligands has been developed. The *Restraints Editor, Especially Ligands* (*REEL*) removes the tedium of manually editing a restraints file by providing a number of commonly performed actions *via* pull-down menus and other interactive features. The effect of changes in the restraints can be immediately reflected in the molecule view to provide user feedback.

A tool that uses many of the features of *eLBOW* to quickly and more easily prepare a protein model for refinement is known as *ReadySet!* The flexibility of the Python interface is exemplified by the use of *Reduce*, *eLBOW* and several smaller portions of the *cctbx* toolkit to add H and/or D atoms to the model, ligands and water and to generate metal-coordination files and geometry restraints for unknown ligands. The files required for covalently bound ligands are also generated.

18.11.7. Integrated structure determination

18.11.7.1. Why automation?

Automation has dramatically changed macromolecular crystallography over the past decade, both by greatly speeding up the process of structure solution, model building and refinement and by bringing the tools for structure determination to a much wider group of scientists. As automation becomes increasingly comprehensive, it will allow users to test many more possibilities for structure determination, will allow improved estimation of uncertainties in the final structures and will allow the determination of ever more complex and difficult structures.

The *PHENIX* environment has been developed with automation as a key and defining feature. Each tool within *PHENIX* can seamlessly and nearly effortlessly be incorporated as part of any other tool or process in *PHENIX*. This means that very complex tasks can be built up from well tested and characterized tools and that tools and higher-level methods can be re-used in many different contexts. With a full automatic regression testing system as an integral part of the *PHENIX* environment, all these tasks and high-level methods are tested daily to ensure the integrity of the entire *PHENIX* system.

18.11.7.2. Automated structure solution

PHENIX has fully integrated structure-solution capability for both experimental phasing [multiwavelength anomalous diffraction (MAD), SAD, multiple isomorphous replacement (MIR) and combinations of these], carried out by *phenix.autosol*, and for molecular replacement, performed by *phenix.automr*. Each of these automated procedures feeds directly into the iterative model building, density modification and refinement of *phenix.autobuild*.

Phenix.autosol is designed to allow complete automation of experimental phasing while allowing a high degree of flexibility for advanced users. Beginning with structure-factor amplitudes and the sequence of the macromolecule, *phenix.autosol* uses *phenix.solve* (Terwilliger & Berendzen, 1999) to scale all data sets, *phenix.xtriage* (Zwart *et al.*, 2005) to analyse the data for twinning and to correct any anisotropy in the data and *phenix.hyss* (Grosse-Kunstleve & Adams, 2003) to find potential heavy-atom or anomalously scattering atoms. *Phenix.autosol* carries out experimental phasing with *phenix.phaser* (McCoy *et al.*, 2004, 2007) or *phenix.solve* (Terwilliger & Berendzen, 1999), density modification with *phenix.resolve* (Terwilliger, 1999) and preliminary model building using the methods in *phenix.autobuild* (Terwilliger, Grosse-Kunstleve, Afonine, Moriarty, Zwart *et al.*, 2008).

A key step in automated structure solution is the identification of which of several possible space-group and heavy-atom or anomalously scattering-atom substructures is correct. *Phenix.autosol* uses a Bayesian scoring algorithm based on analysis of the experimental electron-density maps to identify which substructures lead to the best maps (Terwilliger *et al.*, 2009). The main features of the maps that are used in this evaluation are the skewness of the electron density (non-Gaussian histogram of density with more density in the positive tail than the negative tail) and the correlation of local r.m.s. density (large contiguous regions of high variation where the molecule is located and separate large contiguous regions of low variation where the solvent is located).

Phenix.autosol is highly flexible, allowing any combination of experimental data, such as MAD + SIRAS (single isomorphous replacement with anomalous scattering) or several SAD data sets. Although it is fully automated, the user can control nearly all aspects of the operation of the procedure, including the scoring criteria and decisions about how certain *phenix.autosol* should be that the correct solution is contained in the current lists of solutions.

Phenix.autosol can carry out phasing using a combination of experimental SAD data and molecular-replacement information. If a molecular-replacement model is available, *phenix.autosol* will use *phenix.phaser* (McCoy *et al.*, 2004, 2007) to complete the anomalous substructure iteratively by constructing log-likelihood gradient maps for the anomalous scatterers based on the model of the non-anomalous structure and any anomalous scatterers that have already been found. The anomalous substructure is then used along with the model to calculate phases with *phenix.phaser*.

Phenix.automr carries out automated likelihood-based molecular replacement using *phenix.phaser* (Read, 2001; McCoy *et al.*, 2005, 2007; McCoy, 2007). The procedure is highly automated, allowing several copies of each of several components to be placed in a single run, which can also test different possible choices of space group. If there are alternative choices of model for a component, the molecular-replacement calculation can try each of them in turn or combine them as a statistically weighted ensemble. Although the evaluation of the likelihood targets is slow (Read, 2001), the use of fast approximations for the rotation search (Storoni *et al.*, 2004) and the translation search (McCoy *et al.*, 2005) gives run times that are competitive with traditional Patterson-based methods. Likelihood has been demonstrated to be more sensitive to the correct solution, particularly in difficult cases (Read, 2001). When there are several copies or several components to place, the ability of the likelihood functions to take advantage of preliminary partial solutions can provide a crucial increase in the signal.

18.11.7.3. Iterative model building, density modification and refinement

Phenix.autobuild is a highly integrated and automated procedure for model building and model improvement through iterative model building, density modification and refinement. *Phenix.autobuild* uses *phenix.resolve* (Terwilliger, 2003*a*,*b*) to carry out model building, model extension, model assembly, loop fitting and building outside existing models. It further uses *phenix.resolve* to improve electron-density maps with statistical density modification, including information from the newly built models as well as that obtained from experiment (*e.g. phenix.autosol*), from NCS (Terwilliger, 2002*b*) and from other expected features of electron-density maps such as a flat solvent (Wang, 1985), the presence of secondary-structural features (Terwilliger, 2001) and the presence of local patterns of density characteristic of macromolecules (Terwilliger, 2003*c*). To reduce model bias in the procedure, prime-and-switch phasing can also be used (Terwilliger, 2004). *Phenix.autobuild* uses *phenix.refine* (Afonine *et al.*, 2005) throughout this process to improve the quality of the models that are built.

Phenix.autobuild provides two complementary approaches to model building. For cases in which no model or only a preliminary model has been built, *phenix.autobuild* will construct a new model considering the main chain of any supplied models as potential coordinates. In cases where a nearly final model is available, *phenix.autobuild* can apply a rebuild-in-place approach in which the polypeptide chain is rebuilt a few residues at a time without changing the register or the overall features of the model.

The rebuild-in-place approach in *phenix.autobuild* provides a powerful method for the assessment of uncertainties in an atomic model by repetitive rebuilding of the model using different random seeds for each iteration (Terwilliger, Grosse-Kunstleve *et al.*, 2007). The variability in the coordinates of each atom in the ensemble that is created is a lower bound on the uncertainty of the position of that atom.

18.11.8. Conclusions

Advances in computational methods and algorithms have made it possible to automate the solution of many structures with *PHENIX*. However, many challenges still exist. In particular, the development of automated methods that can be applied at low resolution (worse than 3.0 Å) remains a priority. In this resolution range there are typically too few experimental data to uniquely define the macromolecular structure for automated *ab initio* model building. Thus, methods are required that rely on prior knowledge from existing macromolecular structures to permit productive automated data interpretation. These methods will need to be developed and applied for all stages of structure solution and tightly integrated to maximize the information extracted from the experimental data.

The authors would like to thank the NIH (grant GM063210) and the Phenix Industrial Consortium for support of the Phenix project. This work was supported in part by the US Department of Energy under Contract No. DE-AC02-05CH11231. RJR is supported by a Principal Research Fellowship from the Wellcome Trust (UK). We are grateful to Luc Bourhis and Marat Mustyakimov for contributions to the computational crystallography toolbox and *PHENIX*, Paul Emsley and Bernhard Lohkamp for their guidance interfacing *PHENIX* with *Coot*, the *CCP*4 developers for making the CCP4 library available for inclusion in *PHENIX*, Kevin Cowtan for providing C++ sequence-alignment code, Garib Murshudov and Alexei Vagin for making the monomer library available and our colleagues in the crystallographic community for numerous scientific discussions. We are deeply indebted to David Abrahams for numerous discussions and for accommodating our requirements when developing the Boost.Python library.

This chapter is based on a recent article published in *Acta Crystallographica Section D* (Adams, Afonine *et al.*, 2010).

References

Abrahams, D. & Grosse-Kunstleve, R. W. (2003). *Building hybrid systems with Boost.Python. C/C++ Users J.* **21**, 29–36.

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Cryst.* D**66**, 213–221.

Adams, P. D., Afonine, P., Grosse-Kunstleve, R. W., Read, R. J., Richardson, J. S., Richardson, D. C. & Terwilliger, T. C. (2009). *Recent developments in phasing and structure refinement for macromolecular crystallography. Curr. Opin. Struct. Biol.* **19**, 566–572.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *PHENIX: building new software for automated crystallographic structure determination. Acta Cryst.* D**58**, 1948–1954.

Adams, P. D., Mustyakimov, M., Afonine, P. V. & Langan, P. (2009). *Generalized X-ray and neutron crystallographic analysis: more accurate and complete structures for biological macromolecules. Acta Cryst.* D**65**, 567–573.

Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *The Phenix refinement framework. CCP4 Newsl.* **42**, contribution 8.

Afonine, P. V., Grosse-Kunstleve, R. W., Adams, P. D., Lunin, V. Y. & Urzhumtsev, A. (2007). *On macromolecular refinement at subatomic resolution with interatomic scatterers. Acta Cryst.* D**63**, 1194–1197.

Afonine, P. V., Grosse-Kunstleve, R. W., Chen, V. B., Headd, J. J., Moriarty, N. W., Richardson, J. S., Richardson, D. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2010). *phenix.model_vs_data: a high-level tool for the calculation of crystallographic model and data statistics. J. Appl. Cryst.* **43**, 669–676.

Afonine, P. V., Grosse-Kunstleve, R. W., Urzhumtsev, A. & Adams, P. D. (2009). *Automatic multiple-zone rigid-body refinement with a large convergence radius. J. Appl. Cryst.* **42**, 607–615.

Afonine, P. V., Urzhumtsev, A., Grosse-Kunstleve, R. W. & Adams, P. D. (2010). *Comput. Crystallogr. Newsl.* **1**, 24–31 (http://www.phenix-online.org/newsletter/).

Altona, C. & Sundaralingam, M. (1972). *Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. J. Am. Chem. Soc.* **94**, 8205.

Arendall, W. B., Tempel, W., Richardson, J. S., Zhou, W., Wang, S., Davis, I. W., Liu, Z.-J., Rose, J. P., Carson, W. M., Luo, M., Richardson, D. C. & Wang, B.-C. (2005). *A test of enhancing model accuracy in high-throughput crystallography. J. Struct. Funct. Genomics*, **6**, 1–11.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Chen, V. B., Davis, I. W. & Richardson, D. C. (2009). *KiNG (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. Protein Sci.* **18**, 2403–2409.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res.* **35**, W375–W383.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific LLC, Palo Alto, California, USA. http://www.pymol.org.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics. Acta Cryst.* D**60**, 2126–2132.

Flynn, G. E., Black, K. D., Islas, L. D., Sankaran, B. & Zagotta, W. N. (2007). *Structure and rearrangements in the carboxy-terminal region of SpIH channels. Structure*, **15**, 671–682.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents. J. Am. Chem. Soc.* **118**, 519–528.

Grosse-Kunstleve, R. W. & Adams, P. D. (2003). *Substructure search procedures for macromolecular structures. Acta Cryst.* D**59**, 1966–1973.

Grosse-Kunstleve, R. W., Afonine, P. V., Sauter, N. K. & Adams, P. D. (2005). *cctbx news. Newsletter of the IUCr Commission on Crystallographic Computing*, **5**, 69–91.

Grosse-Kunstleve, R. W. & Brunger, A. T. (1999). *A highly automated heavy-atom search procedure for macromolecular structures. Acta Cryst.* D**55**, 1568–1577.

Grosse-Kunstleve, R. W., Moriarty, N. W. & Adams, P. D. (2009). *Proceedings of ASME 2009 International Design Engineering Technical Conferences*, San Diego, California, 30 August – 2 September 2009. Paper #DETC2009-87737. ASME Press.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework. J. Appl. Cryst.* **35**, 126–136.

Gunčar, G., Klemenčič, I., Turk, B., Turk, V., Karaoglanovic Carmona, A., Juliano, L. & Turk, D. (2000). *Crystal structure of cathepsin X: a flip-flop of the ring of His23 allows carboxy-monopeptidase and carboxy-dipeptidase activity of the protease. Structure*, **8**, 305–313.

Headd, J. J., Immormino, R. M., Keedy, D. A., Emsley, P., Richardson, D. C. & Richardson, J. S. (2009). *Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. J. Struct. Funct. Genomics*, **10**, 83–93.

Kostelecky, B., Saurin, A. T., Purkiss, A., Parker, P. J. & McDonald, N. Q. (2009). *Recognition of an intra-chain tandem 14-3-3 binding site within PKCε. EMBO Rep.* **10**, 983–989.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Intensity statistics in twinned crystals with examples from the PDB. Acta Cryst.* D**62**, 83–95.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Structure validation by Cα geometry: phi, psi and Cbeta deviation. Proteins*, **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *The penultimate rotamer library. Proteins*, **40**, 389–408.

Lunin, V. Yu. (1988). *Use of the information on electron density distribution in macromolecules. Acta Cryst.* A**44**, 144–150.

Lunin, V. Yu. & Skovoroda, T. P. (1995). *R-free likelihood-based estimates of errors for phases calculated from atomic models. Acta Cryst.* A**51**, 880–887.

Lutz, M. & Ascher, D. (1999). *Learning Python.* Sebastopol, California, USA: O'Reilly & Associates.

McCoy, A. J. (2007). *Solving structures of protein complexes by molecular replacement with Phaser. Acta Cryst.* D**63**, 32–41.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Phaser crystallographic software. J. Appl. Cryst.* **40**, 658–674.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Likelihood-enhanced fast translation functions. Acta Cryst.* D**61**, 458–464.

McCoy, A. J., Storoni, L. C. & Read, R. J. (2004). *Simple algorithm for a maximum-likelihood SAD function. Acta Cryst.* D**60**, 1220–1228.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *SnB: crystal structure determination via shake-and-bake. J. Appl. Cryst.* **27**, 613–621.

Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. (2009). *electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Cryst.* D**65**, 1074–1080.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *Breaking good resolutions with ARP/wARP. J. Synchrotron Rad.* **11**, 56–59.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Navaza, J. & Vernoslova, E. (1995). *On the fast translation functions for molecular replacement. Acta Cryst.* A**51**, 445–449.

Padilla, J. E. & Yeates, T. O. (2003). *A statistic for local intensity differences: robustness to anisotropy and pseudo-centering and utility for detecting twinning. Acta Cryst.* D**59**, 1124–1130.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.

Popov, A. N. & Bourenkov, G. P. (2003). *Choice of data-collection parameters based on statistic modelling. Acta Cryst.* D**59**, 1145–1153.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *A graphical user interface to the CCP4 program suite. Acta Cryst.* D**59**, 1131–1137.

Pražnikar, J., Afonine, P. V., Gunčar, G., Adams, P. D. & Turk, D. (2009). *Averaged kick maps: less noise, more signal ... and probably less bias. Acta Cryst.* D**65**, 921–931.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1999). *Detecting outliers in non-redundant diffraction data. Acta Cryst.* D**55**, 1759–1764.

Read, R. J. (2001). *Pushing the boundaries of molecular replacement with maximum likelihood. Acta Cryst.* D**57**, 1373–1382.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J. & Berman, H. M. (2008). *RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution). RNA*, **14**, 465–481.

Schomaker, V. & Trueblood, K. N. (1968). *On the rigid-body motion of molecules in crystals. Acta Cryst.* B**24**, 63–76.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Stewart, J. J. P. (2004). *Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. J. Mol. Model.* **10**, 155–164.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Likelihood-enhanced fast rotation functions. Acta Cryst.* D**60**, 432–438.

Terwilliger, T. C. (1999). *Reciprocal-space solvent flattening. Acta Cryst.* D**55**, 1863–1871.

Terwilliger, T. C. (2001). *Maximum-likelihood density modification using pattern recognition of structural motifs. Acta Cryst.* D**57**, 1755–1762.

Terwilliger, T. C. (2002*a*). *Rapid automatic NCS identification using heavy-atom substructures. Acta Cryst.* D**58**, 2213–2215.

Terwilliger, T. C. (2002*b*). *Statistical density modification with non-crystallographic symmetry. Acta Cryst.* D**58**, 2082–2086.

Terwilliger, T. C. (2003*a*). *Automated main-chain model building by template matching and iterative fragment extension. Acta Cryst.* D**59**, 38–44.

Terwilliger, T. C. (2003*b*). *Automated side-chain model building and sequence assignment by template matching. Acta Cryst.* D**59**, 45–49.

Terwilliger, T. C. (2003*c*). *Statistical density modification using local pattern matching. Acta Cryst.* D**59**, 1688–1701.

Terwilliger, T. C. (2004). *Using prime-and-switch phasing to reduce model bias in molecular replacement. Acta Cryst.* D**60**, 2144–2149.

Terwilliger, T. C., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2007). *Ligand identification using electron-density map correlations. Acta Cryst.* D**63**, 101–107.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). *Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. Acta Cryst.* D**65**, 582–601.

Terwilliger, T. C. & Berendzen, J. (1999). *Automated MAD and MIR structure solution. Acta Cryst.* D**55**, 849–861.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Adams, P. D., Moriarty, N. W., Zwart, P., Read, R. J., Turk, D. & Hung, L.-W. (2007). *Interpretation of ensembles created by multiple iterative rebuilding of macromolecular models. Acta Cryst.* D**63**, 597–610.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Adams, P. D., Read, R. J., Zwart, P. H. & Hung, L.-W. (2008). *Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. Acta Cryst.* D**64**, 515–524.

Terwilliger, T. C., Grosse-Kunstleve, R. W., Afonine, P. V., Moriarty, N. W., Zwart, P. H., Hung, L.-W., Read, R. J. & Adams, P. D. (2008). *Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Cryst.* D**64**, 61–69.

Terwilliger, T. C., Klei, H., Adams, P. D., Moriarty, N. W. & Cohn, J. D. (2006). *Automated ligand fitting by core-fragment fitting and extension into density. Acta Cryst.* D**62**, 915–922.

Tronrud, D. E. (1997). *The TNT refinement package. Methods Enzymol.* **277**, 306–319.

Turk, D. (2007). *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman, pp. 111–122. Heidelberg: Springer.

Urzhumtsev, A. G., Skovoroda, T. P. & Lunin, V. Y. (1996). *A procedure compatible with X-PLOR for the calculation of electron-density maps weighted using an R-free-likelihood approach. J. Appl. Cryst.* **29**, 741–744.

Urzhumtseva, L., Afonine, P. V., Adams, P. D. & Urzhumtsev, A. (2009). *Crystallographic model quality at a glance. Acta Cryst.* D**65**, 297–300.

Wang, B.-C. (1985). *Resolution of phase ambiguity in macromolecular crystallography. Methods Enzymol.* **115**, 90–112.

Wang, X., Kapral, G. J., Murray, L. W., Richardson, D. C., Richardson, J. S. & Snoeyink, J. (2008). *RNABC: forward kinematics to reduce all-atom steric clashes in RNA backbone. J. Math. Biol.* **56**, 253–278.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogens. J. Mol. Biol.* **285**, 1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation*. *J. Mol. Biol.* **285**, 1735–1747.

Zwart, P. H. (2005). *Anomalous signal indicators in protein crystallography*. *Acta Cryst.* D**61**, 1437–1448.

Zwart, P. H., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation*. *CCP4 Newsl.* **43**, contribution 7.

Zwart, P. H., Grosse-Kunstleve, R. W., Lebedev, A. A., Murshudov, G. N. & Adams, P. D. (2008). *Surprises and pitfalls arising from (pseudo)-symmetry*. *Acta Cryst.* D**64**, 99–107.

International Tables for Crystallography (2012). Vol. F, Chapter 18.12, pp. 548–551.

Chapter 18.12. Structure determination in the presence of twinning by merohedry

T. O. Yeates and M. R. Sawaya

18.12.1. Introduction

Twinning by merohedry (see Chapter 11.7) presents challenges for structure determination because the correct intensities are not observed directly. This is a key problem, since the ultimate goal of structure determination is to arrive at agreement between the observed crystallographic quantities and the corresponding values calculated from a model. Common strategies for handling this problem at different stages of the structure-determination process are considered here. In the following discussion, it is helpful to note that, in any given strategy, one of three approaches is generally being taken. Either (i) the effects of twinning are ignored and treated essentially as a source of noise, (ii) the effect twinning has on the observed intensities is reversed – a process called 'detwinning' – in an attempt to recover the true underlying crystallographic intensities, or (iii) the effects of twinning are applied to the model in order to match the twinning present in the observed data. The following discussions apply primarily to twinning by hemihedry, in which just two domain orientations are present. For the treatment of higher forms of twinning (Yu *et al.*, 2009; Barends *et al.*, 2005) or other types of disorder (Wang *et al.*, 2005; Tsai *et al.*, 2009; Hare *et al.*, 2009), the reader is referred to the recent literature.

18.12.2. Detwinning based on observed intensities

Equation (11.7.4.1) in Chapter 11.7 describes how the observed twinned intensities arise from linear combinations of true crystallographic intensities. By inverting that system of equations, one can obtain the crystallographic intensities from the observed intensities, assuming the twin fraction α is known:

$$I(\mathbf{h}_1) = \left[(1-\alpha)I_{\rm obs}(\mathbf{h}_1) - \alpha I_{\rm obs}(\mathbf{h}_2)\right]/(1-2\alpha),$$
$$I(\mathbf{h}_2) = \left[-\alpha I_{\rm obs}(\mathbf{h}_1) + (1-\alpha)I_{\rm obs}(\mathbf{h}_2)\right]/(1-2\alpha),$$

where I are the true crystallographic intensities for two reflections $\mathbf{h}_1$ and $\mathbf{h}_2$ related by the twin operation, and $I_{\rm obs}$ are the corresponding observed intensities. The form of this equation makes it clear that the observed data cannot be detwinned when $\alpha = 1/2$ (*i.e.* perfectly twinned). When $\alpha \neq 1/2$, the crystallographic intensities can be calculated but measurement errors in the observed intensities are magnified according to the term $1/(1 - 2\alpha)$ (Fisher & Sweet, 1980; Grainger, 1969). Therefore, detwinning based on observed intensities alone tends to be problematic when the twin fraction is high. If the twin fraction is low then detwinning is robust, although in those cases the effects of twinning are the least important to consider. As a result, detwinning based only on observed intensities is not practised often. The situation is different once calculated intensities from a model become available, as discussed below.

18.12.3. Molecular replacement with twinning

Even in the case of perfect twinning, it is often possible to arrive at a correct molecular-replacement solution without taking twinning into account, as demonstrated by Redinbo & Yeates (1993). At the rotation-function stage, the effects of twinning are much like the effects of high crystal symmetry; one is attempting to detect the correct orientation of a single search model in the context of noise due to other molecules in distinct orientations within the crystal. In favourable cases such a search can be successful. At the translation-function stage, recognizing the true space-group symmetry is critical. As described in Chapter 11.7, even the correct point-group symmetry can be misclassified in cases of high or perfect twinning. Therefore, it is generally advisable to evaluate potential molecular-replacement solutions in multiple space groups consistent with the data. In practice, twinning is usually ignored throughout the molecular-replacement process (Breyer *et al.*, 1999; Contreras-Martel *et al.*, 2001; Luecke *et al.*, 1998; Redinbo & Yeates, 1993; Larsen *et al.*, 2002; Heffron *et al.*, 2006; Yuan *et al.*, 2003; Lee *et al.*, 2003; Anand *et al.*, 2007), although some workers have detwinned the data and then proceeded with molecular replacement (Rabijns *et al.*, 2001; Taylor *et al.*, 2000). In evaluating potential molecular-replacement solutions, it could be useful to consider the contributions of both domain orientations to the observed intensities; this has generally not been done.

18.12.4. Multiple isomorphous replacement and anomalous phasing with twinning

Strategies for phasing twinned data by multiple isomorphous replacement (MIR) and anomalous-scattering methods have been reviewed (Terwisscha van Scheltinga *et al.*, 2003; Dauter, 2003). To date, several successful solutions of twinned structures by MIR and anomalous phasing approaches have ignored the effects of twinning, at least during the steps of heavy-atom substructure determination and phasing (Rees & Lipscomb, 1980; Mueller, Muller *et al.*, 1999; Mueller, Schübel *et al.*, 1999; MacRae & Doudna, 2007). As noted above, such an approach effectively treats the contributions from the alternate (minor) twin domain as noise. If a correct heavy-atom substructure can be obtained, from Patterson analysis for example, that solution reflects only the heavy-atom contents of one twin domain, presumably the dominant one. Additional peaks not directly accounted for by the identified heavy-atom positions will be present in the native Patterson map in twin-related positions. Adding to the potential confusion, cross-peaks between the heavy-atom positions identified and those related by the twin operation are not present in a Patterson map. Likewise, contributions of the minor twin domain to the measured differences between native and derivative structure-factor amplitudes are ignored when estimating the best phase for the native structure factor. In favourable cases, useful heavy-atom substructures and phases can be obtained despite these serious oversimplifications. In less favourable cases, some workers have overcome twinning in MIR experiments by detwinning the native and derivative data sets, then executing isomorphous replacement with detwinned structure factors (Declercq & Evrard, 2001; Terwisscha van Sheltinga *et al.*, 2001; Hillig & Renault, 2006; Ban *et al.*, 1999). In a few cases, results indicate that detwinning the data produces maps of a quality similar to or worse than the observed twinned data (Rudolph *et al.*, 2003; Yang *et al.*, 2000; Dong *et al.*, 2001).

Copyright © 2012 International Union of Crystallography

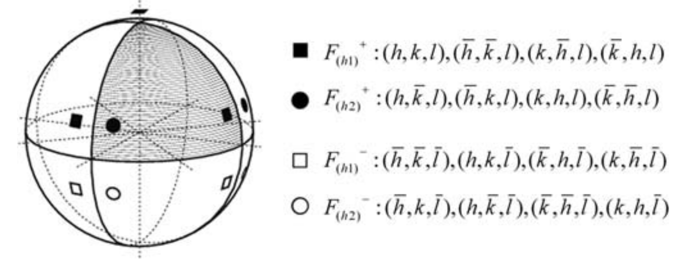

■ $F_{(h1)}^{+}:(h,k,l),(\bar{h},\bar{k},l),(k,\bar{h},l),(\bar{k},h,l)$
● $F_{(h2)}^{+}:(h,\bar{k},l),(\bar{h},k,l),(k,h,l),(\bar{k},\bar{h},l)$
□ $F_{(h1)}^{-}:(\bar{h},\bar{k},\bar{l}),(h,k,\bar{l}),(\bar{k},h,\bar{l}),(k,\bar{h},\bar{l})$
○ $F_{(h2)}^{-}:(\bar{h},k,\bar{l}),(h,\bar{k},\bar{l}),(\bar{k},\bar{h},\bar{l}),(k,h,\bar{l})$

Figure 18.12.4.1

An illustration of how twinning operations can mix together reflections denoted F^+ and F^-. The case of space group $P4$ is illustrated. Filled squares represent a reflection $F^+_{(h1)}$ and its symmetry mates related by crystallographic operations (*i.e.* 90° rotations), while open squares represent their Friedel pairs $F^-_{(h1)}$. Filled circles represent distinct but twin-related reflections, $F^+_{(h2)}$, while open circles represent their Friedel pairs, $F^-_{(h2)}$. On the basis of the asymmetric unit shown (shaded), assuming reflections falling in the designated asymmetric unit are assigned as F^+, the twin operators (hatched lines in the *ab* plane) relate reflection $F^+_{(h1)}$ to reflection $F^-_{(h2)}$ and reflection $F^+_{(h2)}$ to reflection $F^-_{(h1)}$. Detwinning anomalous data – if such a step is deemed appropriate – therefore requires that the symmetry relationships be considered properly rather than relying on F^+ and F^- designations.

A proper treatment of the problem of phasing highly twinned data by MIR has been developed (Yeates & Rees, 1987) which takes into account the simultaneous contributions from both twin domains. The phasing of ordinary MIR data is often presented in terms of a phase circle for the native structure factor. The circle describes the requirement for the sum of the squared real and imaginary components of a structure factor to equal the observed intensity. Additional equations derived from heavy-atom information then lead to the identification of the correct point (or native phase) on this circle. In the case of twinning, taking the case of perfect twinning as an example, what is known is not an individual crystallographic intensity but the sum of two twin-related crystallographic intensities. As a result, each observed intensity leads to an equation specifying the sum of four terms: the squared real and imaginary structure-factor components of two distinct twin-related reflections. This equation is a hypersphere, and equations derived from heavy-atom information can specify the correct point on that hypersphere. This effectively identifies the correct or most probable phases for both twin-related reflections simultaneously. The utility of a proper treatment of twinned MIR data has been demonstrated in synthetic test cases (Yeates & Rees, 1987), but such treatments have not been implemented in standard crystallographic programs. It has been unnecessary in a number of successful MIR cases (Rees & Lipscomb, 1980; Declercq & Evrard, 2001; Terwisscha van Scheltinga *et al.*, 2001; Hillig & Renault, 2006; Ban *et al.*, 1999; Rudolph *et al.*, 2003; Yang *et al.*, 2000; Dong *et al.*, 2001), though there are likely to be situations where such treatment is important for success. The handling of anomalous scattering in the presence of twinning parallels that for MIR. Successful phasing has been achieved using standard programs and approaches while ignoring the presence of twinning, while more exact treatments are possible along lines similar to those advocated for MIR.

When dealing with anomalous data from twinned crystals, sometimes special care is required. It may be counterintuitive, but the twinning operation can mix together reflections carrying an F^+ and an F^- designation in a data set reduced in the correct space-group symmetry. For example, in space group $P4$, one possible choice of asymmetric region in reciprocal space would be the octant with h, k and l all non-negative (ignoring, for simplicity, the redundancy in the principal zones). Reflections recorded in this region (and those related by $P4$ rotational symmetry) would receive F^+ designations, while their Bijvoet pairs would be assigned as F^-. Reflections 123 and 213 are distinct in $P4$, and both would typically be labelled F^+ since they fall in the chosen asymmetric region. These two reflections have Bijvoet pairs $12\bar{3}$ and $21\bar{3}$, respectively, and these would be designated F^-. The twin operation (a dyad about the *xy* diagonal, for example) would mix the 123 and $21\bar{3}$ reflections together and the 213 and $12\bar{3}$ reflections together (Fig. 18.12.4.1). It is therefore important to bear in mind, particularly when contemplating detwinning data having an anomalous signal, that reflections designated F^+ and F^- in the true lower symmetry may be mixed together by the twin operation. This presents no real obstacle as long as the Bijvoet pairs are handled correctly.

Finally, although direct methods of phasing are not discussed in detail here, in cases where the resolution is high enough those methods can also be employed successfully in cases of twinning, as exemplified by the *SHELXD* program (Sheldrick, 2008).

18.12.5. Atomic refinement with twinning

If twinning is present, it must be taken into account during atomic refinement, otherwise the R values remain high. If twinning is present in combination with pseudo-crystallographic noncrystallographic symmetry (NCS), it may be possible to achieve acceptable R values at low resolution but not generally at high resolution. To account for twinning during refinement, the best approach is to use the observed intensity measurements as the target values and incorporate the effects of twinning in the calculated crystallographic quantities. Current refinement programs provide for this (Sheldrick, 2008; Adams *et al.*, 2010; Murshudov *et al.*, 1997; Brünger *et al.*, 1998). Typically, an accurate value for the twin fraction can be determined during the refinement process. Note that it is straightforward to solve for α based on a system of equations of the form given in Chapter 11.7 [equation (11.7.4.1)], where I are calculated from the model and $I_{\rm obs}$ are the observed intensities.

As noted in Chapter 11.7, the presence of twinning is sometimes difficult to detect early in structure determination based on observed intensity data alone, so atomic refinement presents an important opportunity to re-examine the possibility of twinning. As described above, the presence of twinning and an accurate value for the twin fraction can usually be obtained during refinement by comparing observed and calculated intensities. Another method of analysis, which is particularly useful for discriminating between twinning and pseudo-crystallographic NCS, is to examine $R_{\rm twin}$ (*i.e.* the R value between potentially twin-related reflections) and compare the values obtained for $R_{\rm twin}$ when it is evaluated over either observed intensity data or intensities calculated from the model (Fig. 18.12.5.1). This has been referred to as an RvR calculation (Lebedev *et al.*, 2006). The result can be especially instructive when examined as a function of resolution. For cases where twinning occurs together with NCS, the calculated and observed $R_{\rm twin}$ values may be low at low resolution, with $R_{\rm twin}$ over the calculated intensities climbing steeply at higher resolution and $R_{\rm twin}$ over the observed intensities remaining relatively low. When twinning is present in the absence of interfering NCS, the $R_{\rm twin}$ value should be nearly random for the calculated intensities but significantly lower for the observed intensities. An analysis of RvR in a survey of macromolecular structures and diffraction data in the Protein Data Bank (PDB; Berman *et al.*, 2000) has illuminated a number of cases of twinned structures (both merohedral and pseudo-

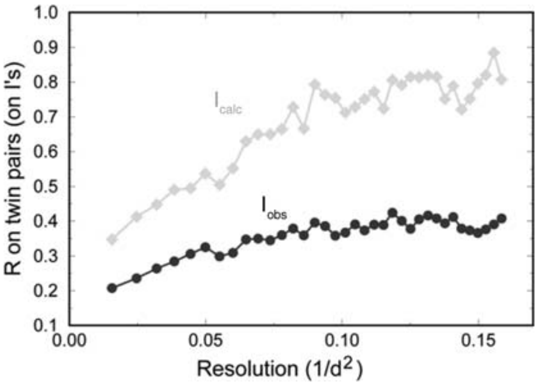


Figure 18.12.5.1
A comparison of calculated *versus* observed $R_{\rm twin}$ values indicates the presence of twinning in a crystal of alanyl-tRNA synthetase (PDB entry 1YFS; Swairjo & Schimmel, 2005). The $I_{\rm obs}$ values continue to obey the NCS/twin-law at high resolution, while the pseudo-symmetry in the $I_{\rm calc}$ values breaks down much more quickly. This particular case depicts pseudo-merohedral twinning in space group $P2_12_12_1$, allowed by nearly equal values of the unit-cell b and c dimensions.

merohedral) where the phenomenon was apparently overlooked during structure determination (Lebedev *et al.*, 2006). This survey emphasizes the importance of checking for twinning during or after atomic refinement.

One cautionary point regarding refinement is that the final R values tend to be artificially low for a twinned refinement as a purely statistical effect. Twinning produces an averaging effect on the intensities as a whole. Therefore, two data sets of twinned intensities (*e.g.* observed and calculated) will tend to differ less from each other than untwinned data sets. It has been suggested that the R values for refining a structure against perfectly twinned data are artificially lower by a factor of approximately $2^{1/2}/2$ (Redinbo & Yeates, 1993). This is in line with the related observation that the expected value of $|L|$ is 1/2 for untwinned data but only 3/8 for perfectly twinned data [see equation (11.7.6.1) in Chapter 11.7 for the definition of $|L|$]. R values must therefore be judged carefully. In particular, the statistical effect means that a drop in the R value when a problem is treated as if it is twinned is generally not good evidence for twinning. Alternative measures of discrepancy could be useful in evaluating the validity of twinned refinement in individual cases. The standard Pearson correlation (*e.g.* between $F_{\rm calc}$ and $F_{\rm obs}$) is likely to be less adversely affected by twinning than the crystallographic R value, but its utility has not been examined in practice.

An important concern when refining twinned structures is how to select test reflections when using the $R_{\rm free}$ value for cross-validation (Brünger, 1992). Selection must be made so that both members of a pair of reflections related by the twin operation are included in the test set. This is relatively simple in cases of twinning by true merohedry, but can be more complicated when dealing with pseudomerohedry.

18.12.6. Detwinning on the basis of model $F_{\rm calc}$ values

As noted above, detwinning based only on observed intensities is usually not advisable. However, once the true crystallographic intensities can be estimated based on model calculations, the situation is improved. Detwinning can be important in two situations. The primary one is to enable electron-density-map calculations. Interpretable electron-density maps require proper estimates for the true crystallographic structure-factor amplitudes ($F_{\rm obs}$). These values can be obtained effectively by detwinning, even in the case of perfect twinning, based on knowledge of the $F_{\rm calc}$ values. At least two different types of equations have been employed, which are different in detail but similar in spirit. We illustrate them in the case of perfect twinning, where the measured $I_{\rm obs} = 1/2[I(\mathbf{h}_1) + I(\mathbf{h}_2)]$ [*i.e.* $I(\mathbf{h}_1) = I(\mathbf{h}_2)$], and the calculated quantities are $I(\mathbf{h}_1)_{\rm calc}$ and $I(\mathbf{h}_2)_{\rm calc}$. In one approach (Wei, 1969), the observed intensity is simply partitioned into estimates for $I(\mathbf{h}_1)$ and $I(\mathbf{h}_2)$ according to the proportion of $I(\mathbf{h}_1)_{\rm calc}$ to $I(\mathbf{h}_2)_{\rm calc}$. This leads to

$$I(\mathbf{h}_1) = \frac{I(\mathbf{h}_1)_{\rm calc}}{\left[I(\mathbf{h}_1)_{\rm calc} + I(\mathbf{h}_2)_{\rm calc}\right]}\,2I_{\rm obs},$$

$$I(\mathbf{h}_2) = \frac{I(\mathbf{h}_2)_{\rm calc}}{\left[I(\mathbf{h}_1)_{\rm calc} + I(\mathbf{h}_2)_{\rm calc}\right]}\,2I_{\rm obs}.$$

Alternatively, one can take a least-squares approach and say that there are two equations for $I(\mathbf{h}_1)$, namely $I(\mathbf{h}_1) = I(\mathbf{h}_1)_{\rm calc}$ and $I(\mathbf{h}_1) = 2I_{\rm obs} - I(\mathbf{h}_2)_{\rm calc}$. Assuming equal estimated errors, the best choice is the average,

$$I(\mathbf{h}_1) = I_{\rm obs} + \frac{\left[I(\mathbf{h}_1)_{\rm calc} - I(\mathbf{h}_2)_{\rm calc}\right]}{2},$$

and likewise for $I(\mathbf{h}_2)$,

$$I(\mathbf{h}_2) = I_{\rm obs} + \frac{\left[I(\mathbf{h}_2)_{\rm calc} - I(\mathbf{h}_1)_{\rm calc}\right]}{2}.$$

Somewhat more complicated expressions arise when $\alpha \neq 1/2$. In either case, one obtains estimates for the true crystallographic intensities. Square roots of these quantities can then be used to calculate electron-density maps. It is critical to note that these values are useful estimates of $F_{\rm obs}$ but they are biased by the current atomic model. Therefore, when calculating electron-density maps using model-based detwinning procedures, it is important to remove model bias locally. As in other kinds of 'omit' maps, atoms in and around a particular region of interest can be removed from the model calculations before applying the detwinning procedure.

Model-based detwinning is also useful in certain atomic refinement situations. If a twinning scenario arises whose treatment is not possible using available refinement programs, model-based detwinning can be used in an iterative fashion to produce structure factors for use as target values in an otherwise ordinary refinement protocol. This strategy was used in the refinement of early cases of protein crystals twinned by hemihedry (Redinbo & Yeates, 1993) and in recent cases of protein crystals twinned by tetartohedry (Yu *et al.*, 2009; Barends *et al.*, 2005; Gayathri *et al.*, 2007).

18.12.7. Summary

Structure determination is often possible even when the only available data are from a perfectly twinned specimen. This is especially true for molecular replacement, but also in favourable cases for MIR and anomalous phasing methods. In general, at the stage where phase determination is being attempted, the available programs attempt to obtain phases (by molecular replacement or heavy-atom methods) while ignoring the presence of twinning. While this has been successful in several cases, there are opportunities for more sophisticated treatments of phase determination in the presence of twinning. Such improvements may make it possible to phase data successfully from a larger number of twinned crystals. At the stage of atomic refinement, the problem of handling twinned structures has been treated rigorously in a number of programs. The proper use of these programs makes it relatively straightforward to refine structures successfully in the presence of twinning.

The authors thank Yingssu Tsai, Regine Herbst-Irmer and George Sheldrick for helpful comments on the manuscript.

References

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Cryst.* D**66**, 213–221.

Anand, K., Schulte, A., Fujinaga, K., Scheffzek, K. & Geyer, M. (2007). *Cyclin box structure of the P-TEFb subunit cyclin T1 derived from a fusion complex with EIAV tat. J. Mol. Biol.* **370**, 826–836.

Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B. & Steitz, T. A. (1999). *Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. Nature (London)*, **400**, 841–847.

Barends, T. R. M., de Jong, R. M., van Straaten, K. E., Thunnissen, A.-M. W. H. & Dijkstra, B. W. (2005). *Escherichia coli MltA: MAD phasing and refinement of a tetartohedrally twinned protein crystal structure. Acta Cryst.* D**61**, 613–621.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Breyer, W. A., Kingston, R. L., Anderson, B. F. & Baker, E. N. (1999). *On the molecular-replacement problem in the presence of merohedral twinning: structure of the N-terminal half-molecule of human lactoferrin. Acta Cryst.* D**55**, 129–138.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Cryst.* D**54**, 905–921.

Contreras-Martel, C., Martinez-Oyanedel, J., Bunster, M., Legrand, P., Piras, C., Vernede, X. & Fontecilla-Camps, J.-C. (2001). *Crystallization and 2.2 Å resolution structure of R-phycoerythrin from Gracilaria chilensis: a case of perfect hemihedral twinning. Acta Cryst.* D**57**, 52–60.

Dauter, Z. (2003). *Twinned crystals and anomalous phasing. Acta Cryst.* D**59**, 2004–2016.

Declercq, J.-P. & Evrard, C. (2001). *A twinned monoclinic crystal form of human peroxiredoxin 5 with eight molecules in the asymmetric unit. Acta Cryst.* D**57**, 1829–1835.

Dong, A., Yoder, J. A., Zhang, X., Zhou, L., Bestor, T. H. & Cheng, X. (2001). *Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. Nucleic Acids Res.* **29**, 439–448.

Fisher, R. G. & Sweet, R. M. (1980). *Treatment of diffraction data from crystals twinned by merohedry. Acta Cryst.* A**36**, 755–760.

Gayathri, P., Banerjee, M., Vijayalakshmi, A., Azeez, S., Balaram, H., Balaram, P. & Murthy, M. R. N. (2007). *Structure of triosephosphate isomerase (TIM) from Methanocaldococcus jannaschii. Acta Cryst.* D**63**, 206–220.

Grainger, C. T. (1969). *Pseudo-merohedral twinning. The treatment of overlapped data. Acta Cryst.* A**25**, 427–434.

Hare, S., Cherepanov, P. & Wang, J. (2009). *Application of general formulas for the correction of a lattice-translocation defect in crystals of a lentiviral integrase in complex with LEDGF. Acta Cryst.* D**65**, 966–973.

Heffron, S. E., Moeller, R. & Jurnak, F. (2006). *Solving the structure of Escherichia coli elongation factor Tu using a twinned data set. Acta Cryst.* D**62**, 433–438.

Hillig, R. C. & Renault, L. (2006). *Detecting and overcoming hemihedral twinning during the MIR structure determination of Rna1p. Acta Cryst.* D**62**, 750–765.

Larsen, N. A., Heine, A., de Prada, P., Redwan, E.-R., Yeates, T. O., Landry, D. W. & Wilson, I. A. (2002). *Structure determination of a cocaine hydrolytic antibody from a pseudomerohedrally twinned crystal. Acta Cryst.* D**58**, 2055–2059.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Intensity statistics in twinned crystals with examples from the PDB. Acta Cryst.* D**62**, 83–95.

Lee, S., Sawaya, M. R. & Eisenberg, D. (2003). *Structure of superoxide dismutase from Pyrobaculum aerophilum presents a challenging case in molecular replacement with multiple molecules, pseudo-symmetry and twinning. Acta Cryst.* D**59**, 2191–2199.

Luecke, H., Richter, H. T. & Lanyi, J. K. (1998). *Proton transfer pathways in bacteriorhodopsin at 2.3 ångström resolution. Science*, **280**, 1934–1937.

MacRae, I. J. & Doudna, J. A. (2007). *An unusual case of pseudo-merohedral twinning in orthorhombic crystals of Dicer. Acta Cryst.* D**63**, 993–999.

Mueller, U., Muller, Y. A., Herbst-Irmer, R., Sprinzl, M. & Heinemann, U. (1999). *Disorder and twin refinement of RNA heptamer double helices. Acta Cryst.* D**55**, 1405–1413.

Mueller, U., Schübel, H., Sprinzl, M. & Heinemann, U. (1999). *Crystal structure of acceptor stem of tRNA(Ala) from Escherichia coli shows unique G.U wobble base pair at 1.16 Å resolution. RNA*, **5**, 670–677.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Rabijns, A., Verboven, C., Novoa de Armas, H., Van Damme, E. J. M., Peumans, W. J. & De Ranter, C. J. (2001). *The crystals of a mannose-specific jacalin-related lectin from Morus nigra are merohedrally twinned. Acta Cryst.* D**57**, 609–611.

Redinbo, M. R. & Yeates, T. O. (1993). *Structure determination of plastocyanin from a specimen with a hemihedral twinning fraction of one-half. Acta Cryst.* D**49**, 375–380.

Rees, D. C. & Lipscomb, W. N. (1980). *Structure of potato inhibitor complex of carboxypeptidase A at 5.5 Å resolution. Proc. Natl Acad. Sci. USA*, **77**, 277–280.

Rudolph, M. G., Kelker, M. S., Schneider, T. R., Yeates, T. O., Oseroff, V., Heidary, D. K., Jennings, P. A. & Wilson, I. A. (2003). *Use of multiple anomalous dispersion to phase highly merohedrally twinned crystals of interleukin-1. Acta Cryst.* D**59**, 290–298.

Sheldrick, G. M. (2008). *A short history of SHELX. Acta Cryst.* A**64**, 112–122.

Swairjo, M. A. & Schimmel, P. R. (2005). *Breaking sieve for steric exclusion of a noncognate amino acid from active site of a tRNA synthetase. Proc. Natl Acad. Sci. USA*, **102**, 988–993.

Taylor, H. O., O'Reilly, M., Leslie, A. G. & Rhodes, D. (2000). *How the multifunctional yeast Rap1p discriminates between DNA target sites: a crystallographic analysis. J. Mol. Biol.* **303**, 693–707.

Terwisscha van Scheltinga, A. C., Valegård, K., Hajdu, J. & Andersson, I. (2003). *MIR phasing using merohedrally twinned crystals. Acta Cryst.* D**59**, 2017–2022.

Terwisscha van Scheltinga, A. C., Valegård, K., Ramaswamy, S., Hajdu, J. & Andersson, I. (2001). *Multiple isomorphous replacement on merohedral twins: structure determination of deacetoxycephalosporin C synthase. Acta Cryst.* D**57**, 1776–1785.

Tsai, Y., Sawaya, M. R. & Yeates, T. O. (2009). *Analysis of lattice-translocation disorder in the layered hexagonal structure of carboxysome shell protein CsoS1C. Acta Cryst.* D**65**, 980–988.

Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Correction of X-ray intensities from single crystals containing lattice-translocation defects. Acta Cryst.* D**61**, 67–74.

Wei, C. H. (1969). *Structural analysis of tetracobalt dodecacarbonyl and tetrarhodium docecacarbonyl. Crystallographic treatments of a disordered structure and a twinned composite. Inorg. Chem.* **8**, 2384–2397.

Yang, F., Dauter, Z. & Wlodawer, A. (2000). *Effects of crystal twinning on the ability to solve a macromolecular structure using multi-wavelength anomalous diffraction. Acta Cryst.* D**56**, 959–964.

Yeates, T. O. & Rees, D. C. (1987). *An isomorphous replacement method for phasing twinned structures. Acta Cryst.* A**43**, 30–36.

Yu, F., Song, A., Xu, C., Sun, L., Li, J., Tang, L., Yu, M., Yeates, T. O., Hu, H. & He, J. (2009). *Determining the DUF55-domain structure of human thymocyte nuclear protein 1 from crystals partially twinned by tetartohedry. Acta Cryst.* D**65**, 212–219.

Yuan, Y.-R., Martsinkevich, O. & Hunt, J. F. (2003). *Structural characterization of an MJ1267 ATP-binding cassette crystal with a complex pattern of twinning caused by promiscuous fiber packing. Acta Cryst.* D**59**, 225–238.

International Tables for Crystallography (2012). Vol. F, Chapter 19.1, pp. 553–556.

PART 19. OTHER EXPERIMENTAL TECHNIQUES

Chapter 19.1. Neutron crystallography: methods and information content

A. A. Kossiakoff

19.1.1. Introduction

Neutron and X-ray crystallography are similar in both their experimental methodologies and in the resulting information content. The principal difference between the two methods is brought about by the characteristic scattering potential of the atom types. The scattering of neutrons by material is not proportional to the atomic number, as is the case in X-ray scattering, but rather depends on the individual nuclear characteristics of each atom type. As seen in Table 19.1.1.1, these characteristics show considerably less deviation and systematic trend among the different atom types. For instance, the heavy atoms in biological material – carbon, oxygen and nitrogen – scatter with about the same magnitude as a lead or uranium atom. In addition, neutrons are scattered by the atomic nuclei, which are essentially point sources, producing diffracted intensity not attenuated by a form-factor fall-off at increasingly higher scattering angles, as is the case in X-ray diffraction (Bacon, 1975).

There are a few atomic nuclei that induce a phase change of 180° in the scattered neutron, which results in negative peaks in a neutron density map. An extremely important example of this is the hydrogen nucleus, with a scattering length of −3.7 f (1 f = 10^{-13} cm). Its isotope, deuterium, on the other hand, scatters to give positive peaks (+6.7 f). The fact that H and D atoms can be so clearly distinguished from one another has very important implications for assessing biophysical parameters, as will be discussed below.

The application of the neutron-diffraction technique, which assigns H-atom positions in proteins and differentiates between H and D atoms, has been mainly focused on structural issues in three research areas: (1) protein reaction mechanisms; (2) protein dynamics; and (3) protein–water interactions (Kossiakoff, 1985, and references therein). It must be pointed out that recent advances in nuclear magnetic resonance have made protein dynamics investigations using H/D exchange procedures much easier than similar experiments by neutron diffraction. Additionally, the advances in ultra-high-resolution X-ray crystallography, which have allowed some level of experimental determination of hydrogen atoms in proteins, have further limited the uniqueness of the neutron method. Nevertheless, a number of important structural issues that are best approached by neutron crystallography remain.

Table 19.1.1.1
Scattering lengths for atom types

Element	Atomic No.	Scattering length (f; 1 f = 10^{-13} cm)
H	1	−3.7
D	1	6.7
C	6	6.6
N	7	9.4
O	8	5.8
Mg	12	5.2
S	16	2.8
Ca	20	4.7
Hg	80	12.7
Pb	82	9.4
U	92	8.5

19.1.2. Diffraction geometries

The general experimental setup involves use of a monochromated beam, employing normal-beam (Caine *et al.*, 1976) or flat-cone geometry (Prince *et al.*, 1978). Both approaches use flat detector surfaces, and thus there is a distortion inherent in all the diffraction phenomena that increases as a function of layer line along the axis of rotation. The extent of this effect can be calculated from the experimental parameters, but, in the case of a linear detector, there is only a moderate amount of flexibility available to make the necessary adjustments. The flat-cone geometry is well suited for a linear detector, since upper-level data fall on an undistorted plane. However, such a scheme requires that the detector be adjusted to different orientations with respect to the spectrometer axis (Prince *et al.*, 1978). In the normal-beam configuration, the crystal is usually mounted on a four-circle goniometer, allowing independent rotations around the φ, χ and ω axes to cover a full sphere of reciprocal space. This method can be efficient when used with a two-dimensional area detector because of the distortion of the diffraction pattern.

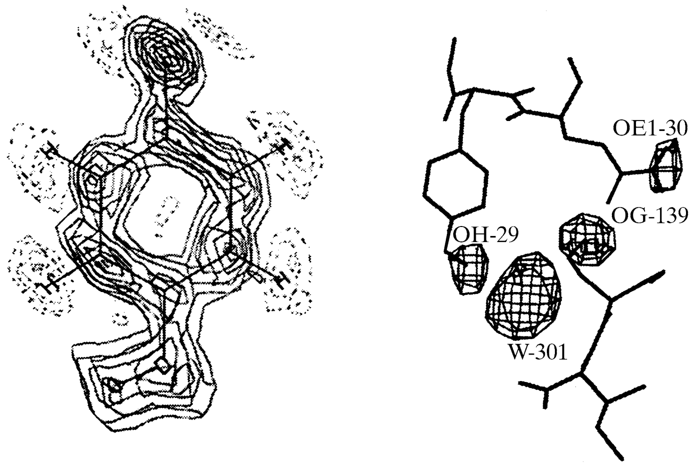
19.1.2.1. *Quasi-Laue diffractometry*

A significant advance in neutron crystallography has been the development and use of modified Laue methods to collect data (Wilkinson & Lehmann, 1991; Wilkinson *et al.*, 1992; Niimura *et al.*, 1997). These methods greatly increase the available neutron flux by using the white neutron spectrum. The full white radiation cannot be used due to very high background scattering and overlap between the diffraction peaks. A reasonable compromise between maximizing intensity while minimizing the experimental problems is to limit the white radiation component to about a 20% wavelength band by employing Ti–Ni multiple-spacing multilayers (Niimura *et al.*, 1997). In practice, the use of the Laue method in X-ray diffraction allows most of the reciprocal space to be recorded in one crystal setting. The quasi-Laue application requires several settings, depending on the neutron intensity distribution $I(\lambda)$ and the crystal symmetry. Data processing can be done using Laue software modified for neutron data.

19.1.3. Neutron density maps – information content

Fig. 19.1.3.1 illustrates several types of structural information derived from neutron density maps. Fig. 19.1.3.1(*a*) shows a well ordered tyrosine ring in the 1.4 Å structure of the protein crambin (Teeter & Kossiakoff, 1984). It can be seen that the ring hydrogen-atom locations are in positions of negative density. These peaks appear to be slightly displaced from their true

Copyright © 2012 International Union of Crystallography



(*a*) (*b*)

Figure 19.1.3.1
Information content in neutron density maps. (*a*) A well ordered tyrosine ring in the 1.4 Å refined structure of crambin (Teeter & Kossiakoff, 1984). (*b*) $D_2O - H_2O$ difference density map of a hydrogen-bonding network in trypsin: Gln30 Oε1–Ser139 D_{γ}, Ser139 O_{γ}–W301, W301–Tyr29 DO. Water density and H/D exchange density shown.

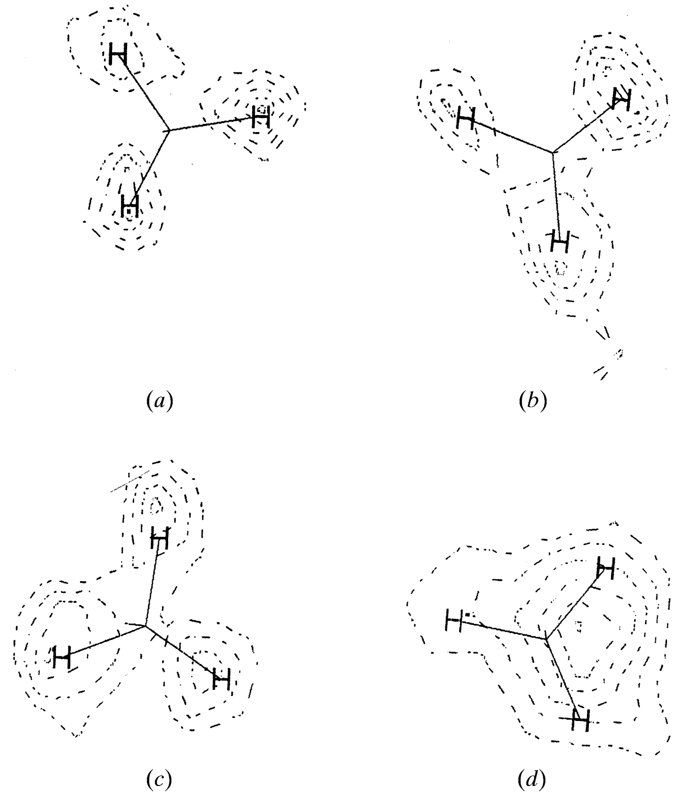
positions, because the map is not at atomic resolution. At 1.4 Å, a portion of the negative peak of the hydrogen overlaps the positive peak of the ring carbon, effectively cancelling density between the atoms and giving the illusion that the peak has been translated. The hydroxyl deuterium orientation is readily determined by its position in positive density.

Use of $D_2O - H_2O$ neutron difference maps provides a high level of stereochemical information (see below) (Kossiakoff *et al.*, 1992; Shpungin & Kossiakoff, 1986). Fig. 19.1.3.1(*b*) displays a network of three hydrogen bonds involving three side-chain types and an occluded water. With knowledge of the heavy atoms alone, it is not possible to define the donor/acceptor character of any of the side chains, because they can act in either capacity, as can the water. The assignments can be made unambiguously from the $D_2O - H_2O$ density, as can the orientation of the water molecule. These maps have allowed detailed analysis of hydroxyl orientations in protein molecules (Kossiakoff *et al.*, 1990; McDowell & Kossiakoff, 1995).

Neutron diffraction is an ideal method for investigating methyl-group conformation, because it allows direct observation of hydrogen-atom positions (Fig. 19.1.3.2) (Kossiakoff & Shteyn, 1984). Although methyl groups in proteins are not held in fixed positions, but spin rapidly around their rotor axes, the time-averaged character of the diffraction experiment establishes the low-energy conformer and the degree of disorder. Accurate methyl-group analysis requires relatively higher resolution (1.5 Å or better) than characterizing other structural features.

19.1.4. Phasing models and evaluation of correctness

Neutron diffraction does not lend itself to the multiple isomorphous phasing approach. This is because the range in atomic scattering power is much narrower than for the X-ray case. There are a few relatively rare isotopes where a significant anomalous effect exists; however, they are not adequate for getting primary phasing information (Schoenborn, 1975). In practice, the initial phasing model has to be derived from the X-ray-determined structure. This is done by applying the appropriate neutron scattering lengths to the refined X-ray coordinates (Norvell & Schoenborn, 1976). Thus, at least in the early stages of analysis, the neutron model relies heavily on the accuracy of the X-ray structure. The importance of an accurate phasing model is borne out by the fact that in several investigations the phasing models were not accurate enough to allow the structure to be refined successfully.



(*a*) (*b*)

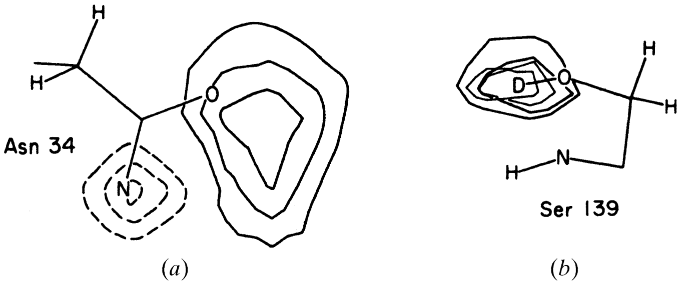
(*c*) (*d*)

Figure 19.1.3.2
Sections of a neutron difference Fourier map showing methyl hydrogen densities for several representative methyl groups. No phasing information about the methyl hydrogens was included in the model; therefore, hydrogens should appear in the difference map at their true positions but at reduced density (~ half weight). The groups shown are: (*a*) Ala24, (*b*) Thr21, (*c*) Thr28, and (*d*) Ala45.

19.1.5. Evaluation of correctness

It is an important first step in the structural analysis to determine the quality of the phases derived from the X-ray structure (Kossiakoff, 1983). Several methods have been used. Using the initial phasing model, the most powerful tests examine an unbiased neutron Fourier map for the appearance of features that are independent of the model. The presence or absence of these features, especially those resulting from the scattering of hydrogen and deuterium atoms, is the most reliable measure of the phasing model. One such test is to evaluate the appearance of the water structure, *i.e.*, the water molecules hydrogen-bonded to the surface of the protein. The water molecules observed in the X-ray analysis are excluded from the neutron-phasing model. The test is applied in cases where the crystals have been soaked in D_2O. The peaks in the neutron density map that correspond to the strongly coordinated water-molecule positions owe their existence solely to the neutron data and phasing model. Even at an early stage, because of the large neutron-scattering potential of D_2O, many of these tightly bound waters found in the X-ray structures should also be observable in the neutron density map.

Another aspect to test phasing reliability is the ability to identify the orientation of side-chain amide groups of asparagine and glutamine. The difference in neutron scattering between O



(*a*) (*b*)

Figure 19.1.5.1

Difference map of Asn34 in the trypsin structure. (*a*) In a protein X-ray analysis, the difference in scattering intensity between O and NH_2 is much too small to be detected. In contrast, the neutron-scattering magnitudes of oxygen and nitrogen (5.8 f *versus* 9.4 f) are quite dissimilar, and there is additional scattering at the nitrogen site from the two bound deuterium atoms. The resulting differential is over 350%, quite large enough to be detected for well ordered side chains. The nitrogen and oxygen positions shown are from the X-ray model. The difference density indicates that the orientation of the nitrogen and oxygen atoms is incorrect and should be rotated by 180° around the C_β–C_γ bond. (*b*) Difference map for Ser139. On well ordered hydroxyl side chains, the orientation of deuterium atoms can sometimes be assigned.

and the two deuteriums and the Nδ2 (5.8 f *versus* 22.6 f) is large enough to be detectable in the Fourier map when these groups are well ordered (Fig. 19.1.5.1). The use of unexchangeable hydrogens for evaluation is considerably more complicated, despite the fact that they constitute about one-half the total number of atoms in the molecule. The difficulty arises from the negative scattering character of the hydrogens, which displaces their apparent positions in the Fourier map from the true positions and, coupled with their short bond lengths, complicates the interpretation of the results. Additionally, it has been shown that small errors in positional and thermal parameters of the parent atoms can further complicate the identification of hydrogen-atom positions (Kossiakoff & Spencer, 1981).

19.1.6. Refinement

The methodologies employed to refine neutron data are essentially the same as those used in most X-ray studies. These include real-space (Hanson & Schoenborn, 1981; Norvell & Schoenborn, 1976; Schoenborn & Diamond, 1976), reciprocal-space (Bentley & Mason, 1980; Phillips, 1984; Wlodawer & Hendrickson, 1982; Wlodawer & Sjolin, 1981) and restrained difference-map refinement (Kossiakoff & Spencer, 1980, 1981). A joint refinement technique in which the neutron and X-ray data are refined simultaneously has been developed (Wlodawer & Hendrickson, 1982). In addition to the normal difficulties encountered in the refinement of any protein structure, there are several that are peculiar to the neutron-diffraction technique. These special problems arise from the close proximity of hydrogen atoms to their parent atoms, coupled with the effects of the negative scattering length of the hydrogen atoms. Potential problems exist when the difference density generated from positional errors of one atom overlaps an adjacent atom site. The situation is further complicated by the fact that, because of its negative scattering length, an error in a hydrogen-atom position is minimized by moving the atom down the gradient, that is, in the opposite direction to that required for correcting parent-atom positions. To evaluate the extent of this problem in refinement, a test was devised using a 2.2 Å data set (Kossiakoff & Spencer, 1981). The coordinates of the protein trypsin were perturbed by a varying, but known, amount from their ideal positions. It was determined that, in general, convergence towards the true coordinate could be obtained when the coordinate errors were less than 0.3 Å; however, if the parent atom (an atom with one or more hydrogens attached to it) was displaced by more than 0.6 Å from its correct position, the effect of neighbouring hydrogens rendered the calculated shifts inaccurate. The results of this study support the observations of other investigators that it is absolutely crucial that the starting phasing model be of high quality, because the range of convergence of a neutron analysis is relatively small.

19.1.7. $D_2O - H_2O$ solvent difference maps

$D_2O - H_2O$ solvent difference maps provide an unbiased method for identifying water molecules and exchangeable hydrogens (Kossiakoff *et al.*, 1992). For several years, the large difference in the scattering characteristics of neutrons by H_2O compared to D_2O has been effectively exploited by using density matching and exchange labelling in small-angle neutron-scattering experiments. This difference can likewise be exploited in neutron protein crystallography to determine the detailed structural characteristics of protein hydration through the calculation of solvent difference maps (Shpungin & Kossiakoff, 1986; Kossiakoff *et al.*, 1992). In practice, such maps are obtained by comparing the changes in diffracted intensities between two sets of data – one obtained from a crystal having H_2O as the major solvent constituent, and a second where D_2O is the solvent medium. To a good approximation, the protein-atom contributions to the scattering intensities in both data sets are equal and cancel, but since H_2O and D_2O have very different scattering properties, their differences are accentuated to reveal an accurate and nearly unbiased representation of the solvent structure.

The features of a solvent difference map of this type are not as affected by errors in the phasing model as conventional difference Fourier maps. In addition, there are refinement procedures that can be applied to them that lead to significant enhancement in signal/noise discrimination. The basic feature of the method is a set of density-modification steps based on the fact that a considerable amount of information about the density distribution of the crystallographic unit cell is known. For instance, it is known that the region of the unit cell occupied by protein atoms should be featureless in solvent maps. It can also be assumed that, as an approximation, solvent regions further than 4 Å from the protein surface have bulk solvent characteristics and can be treated as a constant density region. Combining these two regions gives about 50–60% of the total volume of the unit cell.

Knowledge of the density content of such a large percentage of the unit cell places a strong constraint on the overall character of the Fourier transform, a fact that can be used to improve the quality of the experimentally determined phases

19.1.8. Applications of $D_2O - H_2O$ solvent difference maps

19.1.8.1. *Orientation of water molecules*

In the case of a highly ordered water in a $D_2O - H_2O$ difference map, the oxygen scattering components will cancel (having identical locations and scattering potential), but because H and D have very different scattering properties (H = −3.8 f, D = +6.7 f), large peaks will be found in the map at the D − H positions. Consequently, these ordered waters can usually be oriented with reasonable accuracy in maps better than 2.0 Å resolution (see Fig. 19.1.3.1*b*).

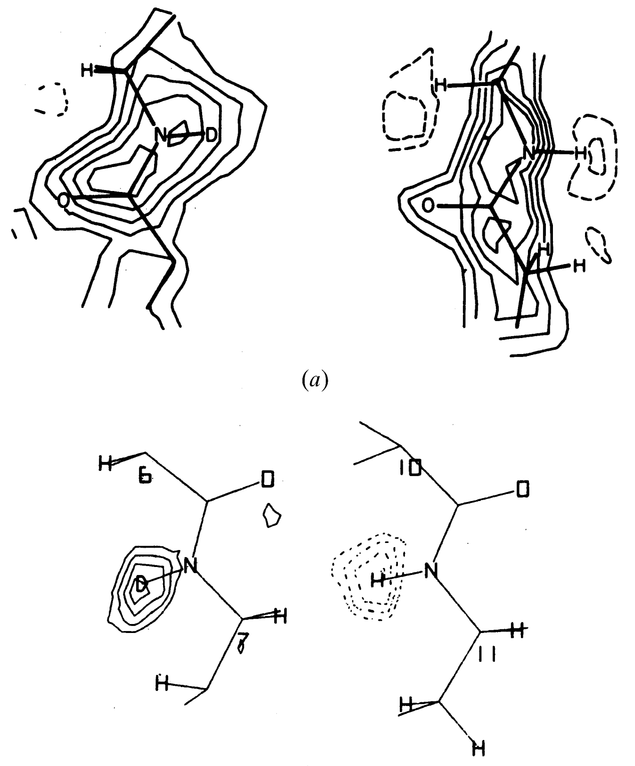


Figure 19.1.8.1
Sections of neutron density maps taken in the plane of the peptide group. (*a*) $2F_o - F_c$ maps showing an example of an exchanged and unexchanged amide peptide group. (*b*) $D_2O - H_2O$ difference density map showing the same.

19.1.8.2. H/D exchange

Fig. 19.1.8.1 shows examples of density around exchanged and unexchanged amide peptide sites. The density in Fig. 19.1.8.1(*a*) is characteristic of $2F_o - F_c$ Fourier maps; the D is in positive density, extending off the peptide nitrogen, and the H is represented by negative density, separated and translated off the nitrogen. Fig. 19.1.8.1(*b*) shows a $D_2O - H_2O$ density map at the same resolution (2.0 Å). Assignment of H/D character is unambiguous, and it is possible to evaluate partial exchange properties more reliably.

References

Bacon, G. E. (1975). *Neutron Diffraction*, pp. 155–188. Oxford: Clarendon Press.

Bentley, G. A. & Mason, S. A. (1980). *Neutron diffraction studies of proteins. Philos. Trans. R. Soc. London Ser. B*, **290**, 505–510.

Caine, J. E., Norvell, J. C. & Schoenborn, B. P. (1976). *Linear position-sensitive counter system for protein crystallography. Brookhaven Symp. Biol.* **27**, 43–50.

Hanson, J. C. & Schoenborn, B. P. (1981). *Real space refinement of neutron diffraction data from sperm whale carbonmonoxymyoglobin. J. Mol. Biol.* **153**, 117–146.

Kossiakoff, A. A. (1983). *Neutron protein crystallography: advances in methods and applications. Annu. Rev. Biophys. Bioeng.* **12**, 259–282.

Kossiakoff, A. A. (1985). *The application of neutron crystallography to the study of dynamic and hydration properties of proteins. Annu. Rev. Biochem.* **54**, 1195–1227.

Kossiakoff, A. A., Shpungin, J. & Sintchak, M. D. (1990). *Hydroxyl hydrogen conformations in trypsin determined by the neutron solvent difference map method: relative importance of steric and electrostatic factors in defining hydrogen-bonding geometries. Proc. Natl Acad. Sci. USA*, **87**, 4468–4472.

Kossiakoff, A. A. & Shteyn, S. (1984). *Effect of protein packing structure on side-chain methyl rotor conformations. Nature (London)*, **311**, 582–583.

Kossiakoff, A. A., Sintchak, M. D., Shpungin, J. & Presta, L. G. (1992). *Analysis of solvent structure in proteins using neutron $D_2O - H_2O$ solvent maps: pattern of primary and secondary hydration of trypsin. Proteins Struct. Funct. Genet.* **12**, 223–236.

Kossiakoff, A. A. & Spencer, S. A. (1980). *Neutron diffraction identifies His 57 as the catalytic base in trypsin. Nature (London)*, **288**, 414–416.

Kossiakoff, A. A. & Spencer, S. A. (1981). *Direct determination of the protonation states of aspartic acid-102 and histidine-57 in the tetrahedral intermediate of the serine proteases: neutron structure of trypsin. Biochemistry*, **20**, 6462–6474.

McDowell, R. S. & Kossiakoff, A. A. (1995). *A comparison of neutron diffraction and molecular dynamics structures: hydroxyl group and water molecular orientations in trypsin. J. Mol. Biol.* **250**, 553–570.

Niimura, N., Minezaki, Y., Nonaka, T., Castagna, J.-C., Cipriani, F., Hoghoj, P., Lehmann, M. S. & Wilkinson, C. (1997). *Neutron Laue diffractometry with an imaging plate provides an effective data collection regime for neutron protein crystallography. Nature Struct. Biol.* **4**, 909–914.

Norvell, J. C. & Schoenborn, B. P. (1976). *Use of the tangent formula for the refinement of neutron protein data. Brookhaven Symp. Biol.* **27**, 1124–1130.

Phillips, S. E. V. (1984). *Hydrogen bonding and exchange in oxymyoglobin*. In *Neutrons in Biology*, edited by B. P. Schoenborn, pp. 305–322. New York: Plenum Press.

Prince, E., Wlodawer, A. & Santoro, A. (1978). *Flat-cone diffractometer utilizing a linear position-sensitive detector. J. Appl. Cryst.* **11**, 173–178.

Schoenborn, B. P. (1975). *Anomalous scattering*. In *Conference on Anomalous Scattering*, edited by S. Rameseshan & S. C. Abrahams, pp. 407–421. Copenhagen: Munksgaard.

Schoenborn, B. P. & Diamond, R. (1976). *Neutron diffraction analysis of metmyoglobin. Brookhaven Symp. Biol.* **27**, 3–11.

Shpungin, J. & Kossiakoff, A. A. (1986). *A method of solvent structure analysis for proteins using $D_2O - H_2O$ neutron difference maps. Methods Enzymol.* **127**, 329–342.

Teeter, M. M. & Kossiakoff, A. A. (1984). *The neutron structure of the hydrophobic plant protein crambin*. In *Neutrons in Biology*, edited by B. P. Schoenborn, pp. 335–348. New York: Plenum Press.

Wilkinson, C., Gabriel, A., Lehmann, M. S., Zemb, T. & Ne, F. (1992). *Image plate neutron detector. Proc. Soc. Photo-Opt. Instrum. Eng.* **1737**, 324–329.

Wilkinson, C. & Lehmann, M. S. (1991). *Quasi-Laue neturon diffractometer. Nucl. Instrum. Methods A*, **310**, 411–415.

Wlodawer, A. & Hendrickson, W. A. (1982). *A procedure for joint refinement of macromolecular structures with X-ray and neutron diffraction data from single crystals. Acta Cryst.* A**38**, 239–247.

Wlodawer, A. & Sjolin, L. (1981). *Orientation of histidine residues in RNase A: neutron diffraction study. Proc. Natl Acad. Sci. USA*, **78**, 2853–2855.

International Tables for Crystallography (2012). Vol. F, Chapter 19.2, pp. 557–562.

Chapter 19.2. Electron diffraction of protein crystals

W. Chiu

19.2.1. Electron scattering

When an electron interacts with a free atom, it is simultaneously attracted to the nucleus because of the nuclear positive charge and repelled by the electrons of the atom. An electron scattering event is a composite of these forces (Hirsch *et al.*, 1977). Mathematically speaking, an electron 'sees' the potential function of the atom, which can be approximated as a 'screened Coulomb potential function'. This function is often referred to as a mass-density function and is analogous to the electron-density function in the case of an X-ray photon, which is scattered only by the electrons of an atom.

Because of the strong interactions between an electron and an atom, the scattering cross section of an atom is much higher for electrons than it is for X-rays. For a 100 keV electron, it is about 10^4 times greater than for an X-ray. For every single electron scattering event of a carbon atom, there is more than a 60% probability that the electron will lose part of its energy, which is called inelastic scattering. The energy lost is primarily in the range 10 to 20 eV, which is sufficient to induce excitation and ionization of the atoms upon irradiation (Isaacson, 1977). This energy transfer to a molecule results in breakage of chemical bonds and mass migration of broken molecular fragments.

19.2.2. The electron microscope

An electron microscope is conceptually analogous to a light microscope. It consists of an electron source, condenser lenses, an objective lens, projector lenses and a camera recorder. Because of the electronegative property of the electron, it is possible to fabricate magnetic and electrostatic lenses to focus electrons to near atomic resolution. The most critical lens in an electron microscope is the objective lens, which forms the first diffraction pattern at its focal plane and the first image at its image plane. An electron diffraction pattern is the same as an X-ray diffraction pattern, containing only the amplitude information of the structure factor, whereas an electron image contains both the amplitude and phase information of the structure factor (Unwin & Henderson, 1975).

The condenser lens is used primarily to control the beam diameter and the flux of the electrons irradiating the specimen, whereas the projector lenses are used to magnify either a diffraction pattern or an image in a broad range. The camera length in an electron microscope is adjustable, ranging from 0.2 to 2.5 m, and allows the recording of diffraction patterns with Bragg spacings from hundreds to a fraction of an ångstrom. The magnification of an image spans from a hundred to a million times. Magnification is not a limiting factor for image resolution, however, and is typically set between 40 and 80 000 times for protein electron crystallography. The most important factors that affect the instrumental resolution are the coherence of the incident electron beam, the chromatic and spherical aberrations of the objective lens, the electrical stability of the electron gun and the objective lens, and the mechanical stability of the specimen stage (Chiu & Glaeser, 1977). In general, almost all modern electron microscopes are capable of resolving a lattice spacing of 2.4 Å in a thin gold crystal. The biological structure resolution in an image of a protein crystal has not reached the instrumental resolution because of many factors related to radiation damage of the specimen. Improvements in experimental and computational methods, however, have made it feasible to image protein crystals beyond 3.7 Å resolution (see below).

The range of electron energy used in an electron microscope is between 20 and 1000 keV, which corresponds to wavelengths of 0.086–0.0087 Å. Any single microscope is built and optimized for a narrow energy range because of the complex design of a highly stable electron gun. The instrument most commonly used for molecular-biology structure research operates in the range 100 to 400 keV. Choosing the most useful electron energy is based on the desired resolution and the specimen thickness. The thicker the specimen, the higher the energy that should be used in order to avoid dynamical scattering effects and to have a sufficient depth of field, so that the electron scattering data can be interpreted with a single scattering theory. Theoretically, the specimen thickness should not exceed about 700 Å if the targeted resolution is 3.5 Å with a 400 kV microscope. Beyond this specimen thickness, the phase error of the structure factors might approach 90° at that resolution. An added advantage of using higher electron energy is to reduce the chromatic aberration effect, resulting in a better-resolved image (Brink & Chiu, 1991).

There are different types of electron emitters, including tungsten filaments, LaB_6 crystals and field emission guns, all of which use different mechanisms to generate the electrons. The field emission gun produces the brightest, most monochromatic beam. The high brightness can allow the electrons to be emitted as from a point source to irradiate the specimen with a highly parallel (*i.e.* a highly spatially coherent) illumination. The benefit of high spatial coherence is the preservation of high-resolution details in the image, even though the defocus of the objective lens is set very high in order to have low-resolution feature contrast (Zhou & Chiu, 1993). Thus, a field emission source is the best choice for high-resolution data collection.

The recording medium of an electron microscope can be an image plate, a slow-scan charge-coupled-device (CCD) camera, or photographic film. Because of their broad dynamic range and high sensitivity, both the CCD camera and the image plate are best suited for recording diffraction patterns (Brink & Chiu, 1994). However, for high-resolution image recording – when the recorded area, pixel resolution, signal-to-noise ratio and the modulation transfer function characteristics must be considered – photographic film is the optimal choice (Sherman *et al.*, 1996).

19.2.3. Data collection

19.2.3.1. Specimen preparation

An electron microscope column is kept at a pressure of $<10^{-6}$ Torr (1 Torr = 133.322 Pa). Because a thin protein crystal loses its crystallinity if dried in a vacuum, its hydration can be maintained by embedding it in a thin layer of vitreous ice, glucose, or other small sugar derivatives (Unwin & Henderson, 1975; Dubochet *et al.*, 1988). The effectiveness of these preser-

Copyright © 2012 International Union of Crystallography

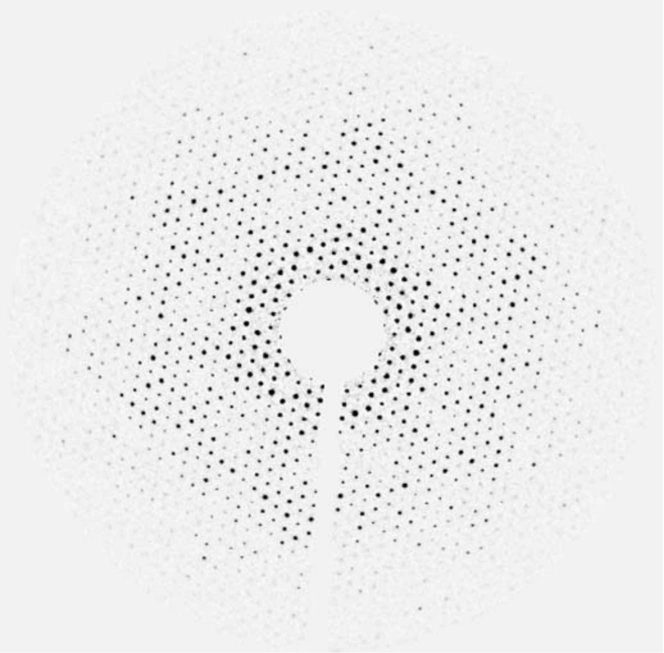


Figure 19.2.3.1

Electron diffraction pattern of trehalose-embedded bacteriorhodopsin, with Bragg reflections extending to 2.5 Å. The unit-cell parameter of this 45 Å-thick membrane protein crystal is 62.5 × 62.5 Å arranged in a $p3$ two-dimensional space group. The raw diffraction pattern was recorded on a Gatan 2k × 2k CCD camera with 300 kV electrons in a JEOL 3000 electron cryomicroscope equipped with a field emission gun and a liquid-helium (4 K) cryoholder. The pattern displayed has been contrast-enhanced using radial background subtraction. A central beam stop was used to prevent saturation of the detector but has blocked off some reflections. $R_{\rm sym}$ for the Friedel-symmetry-related reflections (about 290 pairs) was computed to be about 5%. (Courtesy of Drs Yifan Cheng and Yoshinori Fujiyoshi at Kyoto University.)

vation methods is evidenced by the high-resolution diffraction orders (out to at least 3 Å) from properly embedded protein crystals (Fig. 19.2.3.1). Since the high-resolution reflections come mostly from the protein, their diffraction intensities are largely independent of the embedding medium. However, the low-resolution diffraction intensities can be affected by the embedding medium because different media have different scattering densities relative to the protein. For any new crystal, any of the embedding media mentioned above can be used for high-resolution structural studies.

19.2.3.2. Radiation damage

All protein crystals are prone to radiation damage caused by inelastically scattered electrons (Glaeser, 1971). This physical process is easily seen in the fading of electron diffraction intensities of a protein crystal as the accumulated doses increase. The consequence of damage is a preferential loss of the high-resolution information. Radiation damage is a dose-dependent process and cannot be reduced by adjusting the dose rate (flux) of the irradiating electrons. The strategy used to minimize the damage is to record the diffraction or image data from a specimen area that has not been previously exposed to electrons for purposes of focusing or other adjustments (Unwin & Henderson, 1975). This is called a minimal or low-dose procedure. In addition, keeping a specimen at low temperature (<113 K) allows it to tolerate a higher radiation dose (by a factor of about 4 to 6) before reaching the same extent of damage as at room temperature (Hayward & Glaeser, 1979). It has been shown that damage reduction is minimal below liquid-nitrogen temperature (Chiu *et al.*, 1981). However, there have been some impressive results using the electron cryomicroscope to study membrane protein crystals kept at liquid-helium temperature (4 K) (Kühlbrandt *et al.*, 1994; Kimura *et al.*, 1997; Miyazawa *et al.*, 1999).

19.2.3.3. Other technical factors

In order to record a three-dimensional data set, the crystals have to be tilted to different angles with respect to the direction of the electron beam. In a typical electron microscope, the highest angle to which the specimen stage can be tilted is about ±60°. Consequently, there is a missing set of data beyond the highest tilt angle, which corresponds to no more than 15% of the entire three-dimensional volume. Because of the radiation damage, a single diffraction pattern or a single image per crystal is usually recorded (Henderson & Unwin, 1975). The quality of a crystal is easily judged by its electron diffraction pattern as captured from a CCD camera during data collection. Evaluating the ultimate quality of images, however, takes more time and requires extensive computational analysis.

There are two major technical problems that often limit the data quality, even though a crystal is highly ordered (Henderson & Glaeser, 1985). One is the flatness of the crystal, and the other is the beam-induced movement or charging of the crystal. The effects of both problems become more prominent when the crystals are tilted to high angles. These effects tend to blur the diffraction spots, resulting in loss of high-resolution data (Brink, Sherman *et al.*, 1998). There are many ways to overcome these technical handicaps. For instance, the type of microscope grid chosen or the method of making the carbon support film is critical for reducing the wrinkling of the crystals (Butt *et al.*, 1991; Glaeser, 1992; Booy & Pawley, 1993). The use of a carbon film, which is a good conducting material, to support the protein crystal appears to reduce specimen charging (Brink, Gross *et al.*, 1998). It has been suggested that using a gold-plated objective aperture is effective in reducing specimen charging by generating a stream of secondary electrons to neutralize the positive charges that have built up on the specimen, which thus acts like an aberration-inducing electrostatic lens. Empirically, irradiating the microscope grid before depositing the specimen also reduces the charging (Miyazawa *et al.*, 1999). All these technical problems that can hamper progress in the completion of the structure determination have gradually been identified and resolved. However, more convenient and more robust experimental procedures for reducing these effects further are desirable in order to enhance the efficiency of data collection.

19.2.4. Data processing

19.2.4.1. Data sampling

The principle of three-dimensional reconstruction is based on the central section theorem, which states that the experimental or computed projected diffraction pattern of a three-dimensional object is a plane that intersects the centre of the three-dimensional Fourier space in the direction normal to the direction of the projection (DeRosier & Klug, 1968). Because of the crystallographic symmetry inherent in a protein crystal, only a portion of the entire three-dimensional Fourier space, equivalent to an asymmetric unit of the crystal unit cell, is needed for the reconstruction. The structure factors of a three-dimensional crystal are localized in the three-dimensional reciprocal lattice, whereas the structure factors of a two-dimensional crystal are

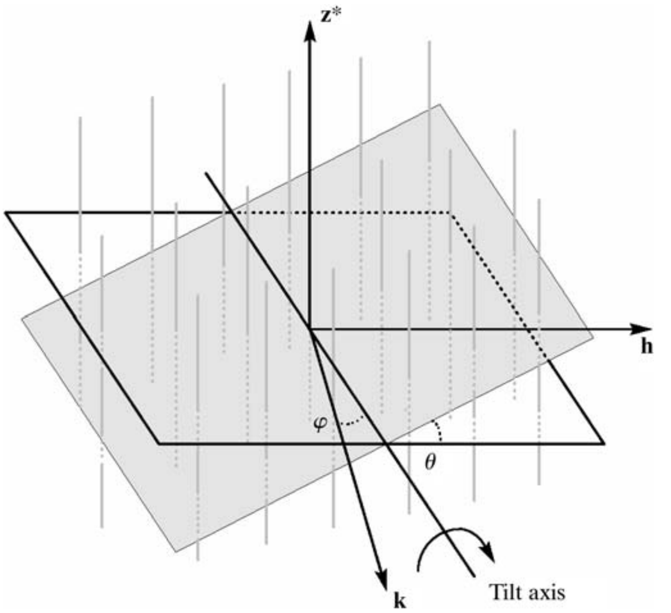


Figure 19.2.4.1
Schematic diagram of data distribution in Fourier space for a two-dimensional crystal. Both the amplitudes and phases of the structure factors are distributed along the lattice lines passing through the (h, k) projection plane. The sectional plane shown denotes data from a crystal tilted at a certain angle (φ), and its tilt axis has an inclined angle (θ) with respect to one of its crystallographic axes. (Courtesy of Dr Jaap Brink at Baylor College of Medicine.)

distributed continuously along the lattice lines, each of which passes through the reciprocal lattice in the zero projection plane (Fig. 19.2.4.1) (Henderson & Unwin, 1975). The assignment of z^* for each observation (h, k, z^*) along the lattice line is determined from the tilt angle and direction of the tilt axis for each image (Shaw & Hills, 1981). In general, the three-dimensional data set is initially built up from low-angle data and is gradually extended to the high-angle data. The angular parameters for each observed reflection are iteratively refined among one another within the whole data set. The required accuracy of these angular determinations depends on the thickness of the crystal and also on the desired resolution (Prasad *et al.*, 1990). For instance, the angular accuracy has to be $< 0.1°$ for 3.5 Å data in a 100 Å-thick crystal. The sampling of the data along the lattice line is generally about $\frac{1}{4}$ the thickness of the crystal (Henderson *et al.*, 1990). The data points are not evenly sampled along the lattice lines; they must be fitted into continuous and smoothly varying functions within the constraint of the crystal thickness. These functions are interpolated onto a periodic lattice, so that its inverse Fourier transform can be computed to reconstruct the three-dimensional mass-density function of the object.

19.2.4.2. Amplitudes and phases

An electron-microscope image contains both the amplitudes and phases of the structure factors. The basic premise of the current image-reconstruction scheme assumes that the image intensity can be related to the structure factor linearly and can be retrieved by the Fourier transform of the image intensities. However, the structure factors, $F(S)$, are influenced by several instrumental factors, as shown in equations (19.2.4.1)–(19.2.4.3) below, whose parameters need to be determined for each image.

$$F^2_{\rm obs}(S) = [F(S){\rm CTF}(S)E(S)]^2 + N^2(S), \qquad (19.2.4.1)$$

$${\rm CTF}(S) = -\{(1-Q^2)^{1/2}\sin[\gamma(S)] + Q\cos[\gamma(S)]\} \text{ and } \qquad (19.2.4.2)$$

$$\gamma(S) = \pi[(-C_s\lambda^3S^4)/2 + \Delta z\lambda S^2], \qquad (19.2.4.3)$$

where $F_{\rm obs}$ is the structure factor computed from the electron cryomicroscopic images, F is the true structure factor, CTF is the contrast-transfer function, E is the product of many decay functions due to the electron optics and specimen movement, N is the background noise contributed by a variety of physcial effects, S is the spatial frequency, Q is the fraction of amplitude contrast, C_s is the spherical aberration coefficient of the objective lens, λ is the wavelength and Δz is the image defocus.

In practice, it is tedious to determine all the parameters in these equations from images in order to make corrections to the amplitudes of the structure factors. In the case of crystals, the amplitudes of the structure factors can simply be obtained directly from the electron diffraction intensities, which are free from any of the above factors (Unwin & Henderson, 1975). The computational procedure used to calculate the diffraction spot intensities is similar to that used to measure an X-ray diffraction pattern (Baldwin & Henderson, 1984; Brink & Wei Tam, 1996). The quality of the diffraction intensity measurement is evaluated from the value of $R_{\rm sym}$ for Friedel-related reflections. The best data have $R_{\rm sym}$ less than 0.04. The consistency of the diffraction intensities among different patterns from different crystals is judged from $R_{\rm merge}$, which is generally 0.15–0.25 (Kimura *et al.*, 1997; Nogales *et al.*, 1998). Fig. 19.2.4.2 is an example of the diffraction intensity for two lattice lines computed from bacteriorhodopsin crystals. The phases of the structure factors are computed from images.

In addition to the instrumental factors given in equations (19.2.4.1)–(19.2.4.3), however, images are generally imperfect because of bending of the crystal, specimen preparation, or magnification variations across an image. The consequence of these imperfections is a reduction of the signal-to-noise ratio in high-resolution reflections. A computational procedure called ‘unbending’ has been devised, which in effect fixes the image imperfection by finding the unit-cell deviation vectors and straightening them by interpolation (Henderson *et al.*, 1986). The effect of the instrumental factors is the modulation of the phases by the oscillating function CTF(S), as shown in equations (19.2.4.1)–(19.2.4.3). The result is that the phases flip by π at different frequencies, depending on the defocus setting (Erickson & Klug, 1970). In addition, there is a phase shift caused by a combination of factors, including lens astigmatism, beam tilt and specimen height variation in a tilted position. All of these factors have to be corrected for each micrograph before merging the phases of the reflections from different micrographs to a common phase origin. The determination of the phase origin is performed by phase residual difference minimization or correlation matches among different micrographs (Amos *et al.*, 1982; Thomas & Schmid, 1995). Intensities and corresponding phases of two lattice lines are shown in Fig. 19.2.4.2. The fitted curves show the matches among the data points, each of which is from a different image or from different symmetry-related reflections from the same image.

In electron crystallography, the correctness of the phases can be evaluated by the self-consistency of the merged data sets and also by the phase residual difference of the symmetry-related reflections according to the two-dimensional plane-group symmetry. For two-dimensional crystals, there are only 17 possible plane groups (Amos *et al.*, 1982). As in the case of a three-dimensional crystal, the plane group is determined from the symmetry of the phases, the unit-cell parameters and the pattern of forbidden reflections. The plane-group assignment can be confirmed by the phase equivalence of symmetry-related reflections. Furthermore, the reliability of the map can be judged

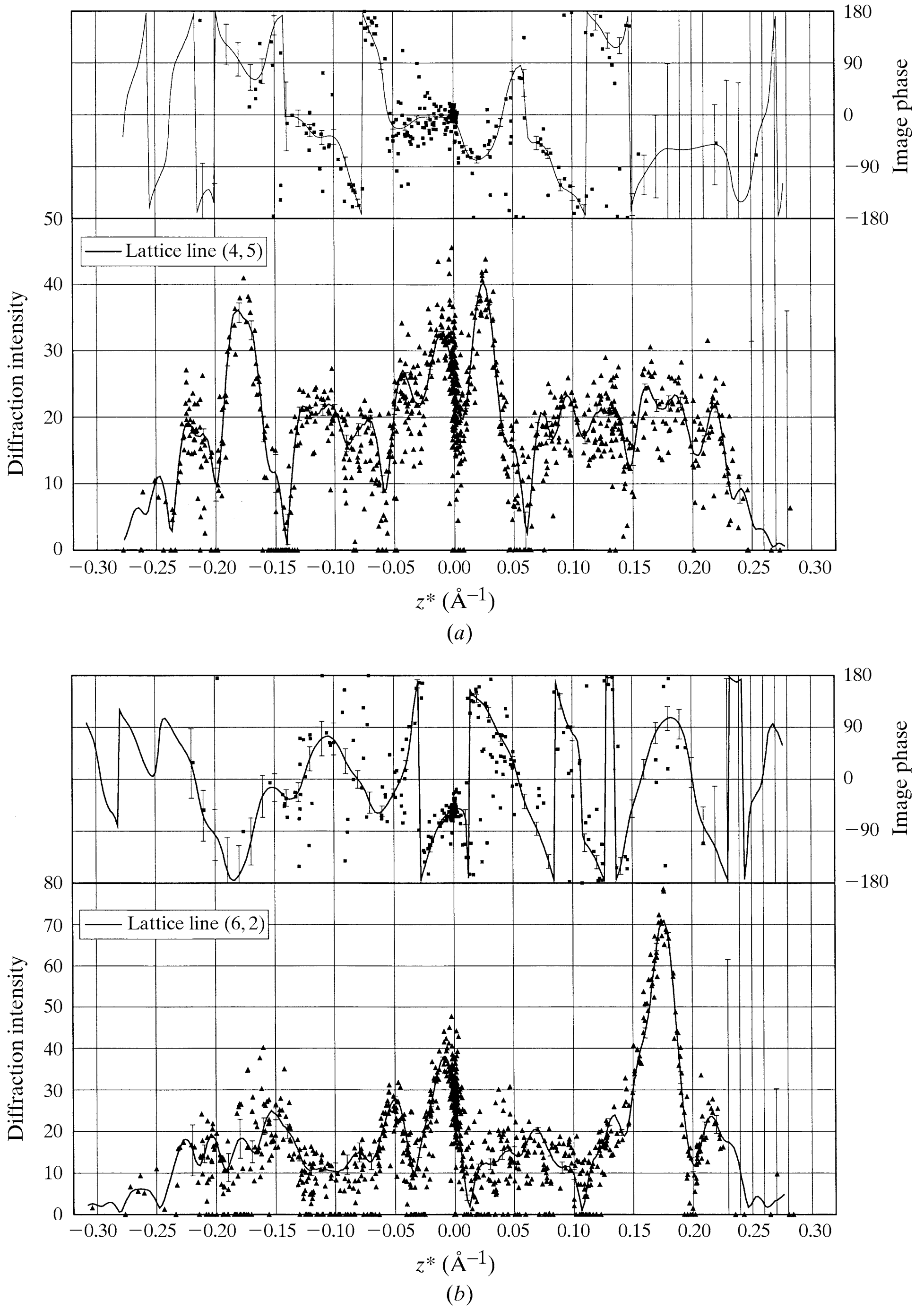


Figure 19.2.4.2

Experimental intensities from electron diffraction patterns and phases from images of bacteriorhodopsin, recorded from tilted crystals in an electron cryomicroscope. Fitted curves for two representative lattice lines are shown: (*a*) (4, 5, z^*) and (*b*) (6, 2, z^*) (Courtesy of Drs Terushisa Hirai and Yoshinori Fujiyoshi at Kyoto University.)

by the figure of merit of the phases, computed from the phase probability distribution function of the observed reflections.

19.2.4.3. 3D map

The three-dimensional (3D) map is computed from the amplitudes and phases at the resolution defined by the data (Henderson & Unwin, 1975). The resolution reported for the structure is defined by the observed reflections in the images. Owing to the missing data at high tilt angles, the reconstruction normally has a lower resolution in the direction of the electron beam than in the direction normal to it. As a result, many of the initial low-resolution structures appear stretched out along the vertical direction. The interpretation of the 3D map derived from electron crystallography is similar to that of X-ray crystallography. Often, the initial map is reported at about 7 Å, where some of the α-helices can be interpreted. With an improved map of about 3.5 Å, the polypeptide backbone is traced and some of the bulky side chains are recognized. Fig. 19.2.4.3 shows a chain tracing of a tubulin crystal (Nogales *et al.*, 1998).

19.2.4.4. Refinement

In order to arrive at a correct mechanistic model for the protein, an accurate atomic structure is needed. So far, in electron crystallography only bacteriorhodopsin has been refined (Grigorieff *et al.*, 1996). A common criterion used in X-ray crystallography to evaluate the progress of refinement is based on

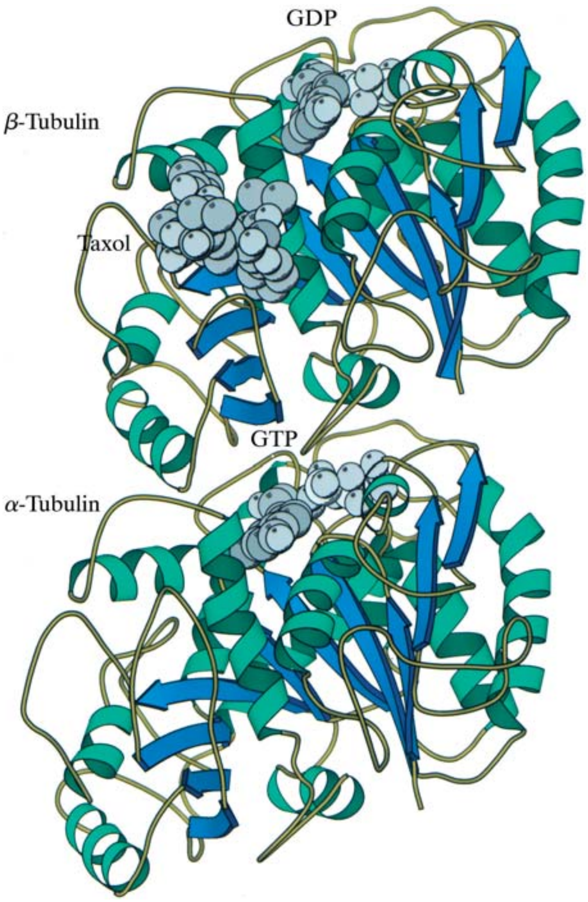


Figure 19.2.4.3
Ribbon diagram of a tubulin dimer, whose structure has been solved to 3.7 Å resolution. GTP, GDP and taxol are shown as CPK models, with GDP at the top, bound to β tubulin, and GTP in the middle, bound to the α subunit. This is the view as seen from inside a microtubule, with the plus end at the top. (Courtesy of Drs Eva Nogales and Kenneth Downing at the Lawrence Berkeley National Laboratory, University of California, Berkeley.)

the free R factor, which measures the agreement between the model and a part of the experimental data not included in the refinement process. In electron crystallography, the phases are measured independently from images and hence are not refined. Therefore, they can be used as a 'free phase residual,' which is analogous to the free R factor, to assess the progress of refinement. The refined structure would result in improved peptide geometry, increased accuracy of the coordinates of the polypeptide backbone and of the amino-acid side chain residues, and improved temperature factors of the residues.

19.2.5. Future development

Electron crystallography has proven to be a high-resolution structural tool for two-dimensional protein crystals, to the point where the polypeptide backbone can be traced and atomic coordinates derived. Needless to say, there is still much to be learned about how to make highly ordered two-dimensional crystals from either membrane or soluble proteins. Research in this direction is critical for the growth of electron crystallography. Recent results have promoted optimism; there has been an increase in the number of membrane proteins crystallized into two-dimensional arrays from which at least 6 to 8 Å structures can be obtained (Walz *et al.*, 1997; Auer *et al.*, 1998; Zhang *et al.*, 1998; Unger *et al.*, 1999).

In the most recent high-resolution structural study of tubulin, a 3.7 Å map was obtained from 100 electron diffraction patterns and 150 electron images. Effectively, this structure was the result of a computational average of about one million tubulin dimers. It took six years to determine the structure from the time when the first high-resolution crystal structure was reported (Downing & Jontes, 1992; Nogales *et al.*, 1998). All the experimental and computational procedures were basically the same as those developed for bacteriorhodopsin (Henderson *et al.*, 1990). An obvious future development in protein electron crystallography would be aimed at improving the throughput of the structural determination. This entails a search for better solutions to some of the technical problems mentioned above as well as the introduction of automation in both data collection and processing.

Finally, another potentially exciting aspect of electron crystallography is the ability to detect charged residues from the high scattering differences between neutral and charged atoms. This physical property may make electron crystallography a unique method for detecting the ionization state of the amino-acid residues in proteins (Mitsuoka *et al.*, 1999). Furthermore, there is also a good prospect of extending the structure close to 2 Å resolution, as the next generation of electron cryomicroscope will be equipped with a field emission gun operated at 300 keV, a liquid helium cryo-specimen stage and an energy filter. This combination of instrumental features is likely to bring electron crystallography a step closer to its ultimate potential for structural biology research at the atomic level.

Research has been supported by RR002250. I thank Drs Jacob Brink, Michael Schmid, Karou Mitsuoka and Ken Downing for helpful comments on the manuscript.

References

Amos, L. A., Henderson, R. & Unwin, P. N. (1982). *Three-dimensional structure determination by electron microscopy of two-dimensional crystals*. *Prog. Biophys. Mol. Biol.* **39**, 183–231.

Auer, M., Scarborough, G. A. & Kühlbrandt, W. (1998). *Three-dimensional map of the plasma membrane H$^+$-ATPase in the open conformation*. *Nature (London)*, **392**, 840–843.

Baldwin, J. & Henderson, R. (1984). *Measurement and evaluation of electron diffraction patterns from two-dimensional crystals*. *Ultramicroscopy*, **14**, 319–336.

Booy, F. P. & Pawley, J. B. (1993). *Cryo-crinkling: what happens to carbon films on copper grids at low temperature*. *Ultramicroscopy*, **48**, 273–280.

Brink, J. & Chiu, W. (1991). *Contrast analysis of cryo-images of n-paraffin recorded at 400 kV out to 2.1 Å resolution*. *J. Microsc.* **161**, 279–295.

Brink, J. & Chiu, W. (1994). *Applications of a slow-scan CCD camera in protein electron crystallography*. *J. Struct. Biol.* **113**, 23–34.

Brink, J., Gross, H., Tittmann, P., Sherman, M. B. & Chiu, W. (1998). *Reduction of charging in protein electron cryomicroscopy*. *J. Microsc.* **191**, 67–73.

Brink, J., Sherman, M. B., Berriman, J. & Chiu, W. (1998). *Evaluation of charging on macromolecules in electron cryomicroscopy*. *Ultramicroscopy*, **72**, 41–52.

Brink, J. & Wei Tam, M. (1996). *Processing of electron diffraction patterns acquired on a slow-scan CCD camera*. *J. Struct. Biol.* **116**, 144–149.

Butt, H. J., Wang, D. N., Hansma, P. K. & Kühlbrandt, W. (1991). *Effect of surface roughness of carbon support films on high-resolution electron diffraction of two-dimensional protein crystals*. *Ultramicroscopy*, **36**, 307–318.

Chiu, W. & Glaeser, R. M. (1977). *Factors affecting high resolution fixed-beam transmission electron microscopy*. *Ultramicroscopy*, **2**, 207–217.

Chiu, W., Knapek, E., Jeng, T. W. & Dietrick, I. (1981). *Electron radiation damage of a thin protein crystal at 4 K*. *Ultramicroscopy*, **6**, 291–296.

DeRosier, D. J. & Klug, A. (1968). *Reconstruction of three-dimensional structures from electron micrographs*. *Nature (London)*, **217**, 130–134.

Downing, K. H. & Jontes, J. (1992). *Projection map of tubulin in zinc-induced sheets at 4 Å resolution*. *J. Struct. Biol.* **109**, 152–159.

Dubochet, J., Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A. W. & Schultz, P. (1988). *Cryo-electron microscopy of vitrified specimens*. *Q. Rev. Biophys.* **21**, 129–228.

Erickson, H. P. & Klug, A. (1970). *The Fourier transform of an electron micrograph: effects of defocusing and aberrations, and implications for the use of underfocus contrast enhancement*. *Philos. Trans. R. Soc. London Ser. B*, **261**, 105–118.

Glaeser, R. M. (1971). *Limitations to significant information in biological electron microscopy as a result of radiation damage*. *J. Ultrastruct. Res.* **36**, 466–482.

Glaeser, R. M. (1992). *Specimen flatness of thin crystalline arrays: influence of the substrate*. *Ultramicroscopy*, **46**, 33–43.

Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. (1996). *Electron-crystallographic refinement of the structure of bacteriorhodopsin*. *J. Mol. Biol.* **259**, 393–421.

Hayward, S. B. & Glaeser, R. M. (1979). *Radiation damage of purple membrane at low temperature*. *Ultramicroscopy*, **4**, 201–210.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). *Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy*. *J. Mol. Biol.* **213**, 899–929.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. (1986). *Structure of purple membrane from Halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution*. *Ultramicroscopy*, **19**, 147–178.

Henderson, R. & Glaeser, R. M. (1985). *Quantitative analysis of image contrast in electron micrographs of beam-sensitive crystals*. *Ultramicroscopy*, **16**, 139–150.

Henderson, R. & Unwin, P. N. (1975). *Three-dimensional model of purple membrane obtained by electron microscopy*. *Nature (London)*, **257**, 28–32.

Hirsch, P., Howie, A., Nicholson, R. B., Pashley, D. W. & Whelan, M. J. (1977). *Electron Microscopy of Thin Crystals*. Huntington: Robert K. Krieger Publishing Co.

Isaacson, M. S. (1977). *Specimen damage in the electron microscope*. In *Principles and Techniques of Electron Microscopy. Biological Applications*, edited by M. A. Hyatt, pp. 1–78. New York: Van Nostrand Reinhold Co.

Kimura, Y., Vassylyev, D. G., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T. & Fujiyoshi, Y. (1997). *Surface of bacteriorhodopsin revealed by high-resolution electron crystallography*. *Nature (London)*, **389**, 206–211.

Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. (1994). *Atomic model of plant light-harvesting complex by electron crystallography*. *Nature (London)*, **367**, 614–621.

Mitsuoka, K., Hirai, T., Murata, K., Miyazawa, A., Kidera, A., Kimura, Y. & Fujiyoshi, Y. (1999). *The structure of bacteriorhodopsin at 3.0 Å resolution based on electron crystallography: implication of the charge distribution*. *J. Mol. Biol.* **286**, 861–882.

Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. (1999). *Nicotinic acetylcholine receptor at 4.6 Å resolution: transverse tunnels in the channel wall*. *J. Mol. Biol.* **288**, 765–786.

Nogales, E., Wolf, S. G. & Downing, K. H. (1998). *Structure of the alpha beta tubulin dimer by electron crystallography*. *Nature (London)*, **391**, 199–203.

Prasad, B. V., Degn, L. L., Jeng, T. W. & Chiu, W. (1990). *Estimation of allowable errors for tilt parameter determination in protein electron crystallography*. *Ultramicroscopy*, **33**, 281–285.

Shaw, P. J. & Hills, G. J. (1981). *Tilted specimen in the electron microscope: a simple specimen holder and the calculation of tilt angles for crystalline specimens*. *Micron*, **12**, 279–282.

Sherman, M. B., Brink, J. & Chiu, W. (1996). *Performance of a slow-scan CCD camera for macromolecular imaging in a 400 kV electron cryomicroscope*. *Micron*, **27**, 129–139.

Thomas, I. M. & Schmid, M. F. (1995). *A cross-correlation method for merging electron crystallographic image data*. *J. Micros. Soc. Am.* **1**, 167–173.

Unger, V. M., Kumar, N. M., Gilula, N. B. & Yeager, M. (1999). *Three-dimensional structure of a recombinant gap junction membrane channel*. *Science*, **283**, 1176–1180.

Unwin, P. N. & Henderson, R. (1975). *Molecular structure determination by electron microscopy of unstained crystalline specimens*. *J. Mol. Biol.* **94**, 425–440.

Walz, T., Hirai, T., Murata, K., Heymann, J. B., Mitsuoka, K., Fujiyoshi, Y., Smith, B. L., Agre, P. & Engel, A. (1997). *The three-dimensional structure of aquaporin-1*. *Nature (London)*, **387**, 624–627.

Zhang, P., Toyoshima, C., Yonekura, K., Green, N. M. & Stokes, D. L. (1998). *Structure of the calcium pump from sarcoplasmic reticulum at 8-Å resolution*. *Nature (London)*, **392**, 835–839.

Zhou, Z. H. & Chiu, W. (1993). *Prospects for using an IVEM with a FEG for imaging macromolecules towards atomic resolution*. *Ultramicroscopy*, **49**, 407–416.

International Tables for Crystallography (2012). Vol. F, Chapter 19.3, pp. 563–574.

Chapter 19.3. Small-angle X-ray scattering

H. Tsuruta and J. E. Johnson

19.3.1. Introduction

Mechanistic biology is frequently confronted with an experimental paradox. High-resolution structures are required to develop a chemical description of macromolecular interactions, but the processes themselves are dynamic and not amenable to high-resolution methods. Chemists and molecular biologists have been successful in generating homogeneous components and entire complex systems that are amenable to crystallization and high-resolution analysis. If the dynamics of the systems are small, as in the case of some enzymes and electron-transfer reactions, time-resolved studies can be performed in the context of the crystal lattice using Laue diffraction (Moffat, 1997; Genick *et al.*, 1997; Srajer *et al.*, 1996). If the dynamic features of the biological reactions are large, the motility must be studied in solution. During the last decade, small-angle X-ray scattering (SAXS) has emerged as an important method for studying large-scale dynamic processes, ranging from protein folding to virus particle polymorphism. The renaissance of this method has resulted from a variety of advances in molecular biology and X-ray instrumentation, and these have dramatically increased the information content of the derived results. Modern synchrotron X-ray sources and advanced detector systems have lead to higher-resolution data in both the spatial and time domains. Computational analyses of the data have improved dramatically, resulting, in some favourable cases, in *de novo* three-dimensional density functions from SAXS data that are comparable to a 20 Å-resolution electron-microscopy reconstruction. Model-based approaches to interpreting SAXS data have expanded their usefulness to the validation of mechanistic hypotheses involving movement or association of components independently determined at high resolution. Finally, SAXS studies have benefited from the same molecular-biology advances that provide large quantities of homogeneous material for crystallographic and NMR studies. The purpose of this chapter is to address practical aspects of SAXS as they relate to and complement macromolecular crystallography.

19.3.2. Small-angle single-crystal X-ray diffraction studies

We begin by briefly describing an obvious connection between small-angle solution X-ray scattering and single-crystal X-ray scattering, *i.e.* small-angle single-crystal diffraction (Tsuruta *et al.*, 1998; Miller *et al.*, 1999). The majority of macromolecular single-crystal studies ignore data at resolutions lower than 20 Å. Measuring these data accurately is generally difficult when collecting high-resolution data, because they are orders of magnitude stronger in intensity and frequently saturate the detector pixels or are intentionally blocked from reaching the detector by a large beam stop. It is technically challenging to record low-resolution data accurately, because they fall very close to the primary beam, where a variety of parasitic scattering effects occur. These data, however, contain information about the structure that may not be determined from only the high-resolution data. Low-resolution (∞–15 Å) data are sensitive to structures that are organized over large distances and that contribute primarily to the low-frequency terms in a Fourier series. In contrast, portions of the structure that contribute to high-frequency terms must be in precisely the same position in all the unit cells within the crystal. Two examples illustrate the importance of low-resolution terms.

Assume that an exposed loop on the surface of a protein is highly mobile. The scattering-factor curves for atoms in this loop will be highly attenuated by the large Debye–Waller factor. If the temperature factors are high enough, these atoms will contribute nothing to the high-resolution data. In contrast, these atoms will contribute strongly to the low-resolution data. If the low-resolution terms are included in the Fourier series, the atom positions will not appear as individual atoms; instead, an envelope of density describing the statistically distributed positions of these atoms will be visible, and this can aid significantly in the modelling of the loop.

A second example is the nucleic acid within a spherical virus. The nucleotides generally do not display the symmetry of the icosahedral capsid, although in some cases segments of RNA or DNA clearly interact with the protein and are therefore visible at high resolution. The RNA density is not usually visible if only higher-resolution data are used, and the virus particle appears empty. The reason for this is that the RNA is best described as a

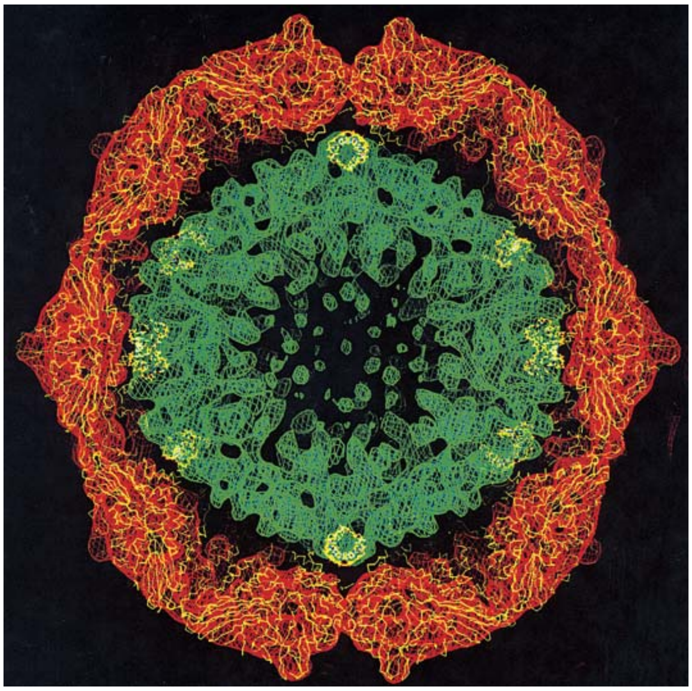


Figure 19.3.2.1
A cross section of the electron-density map of sFHV at 14 Å resolution. The density shown in red corresponds to the protein capsid, while density in green corresponds to the bulk RNA or dynamic protein segments. The Cα backbone of the atomic model of the protein capsid seen in the high-resolution structures of sFHV is shown as yellow traces, and the model of the ordered pieces of duplex RNA is shown as a stick model. The high-resolution protein model fits the electron-density map very well. The ordered RNA model is buried in the RNA density in green.

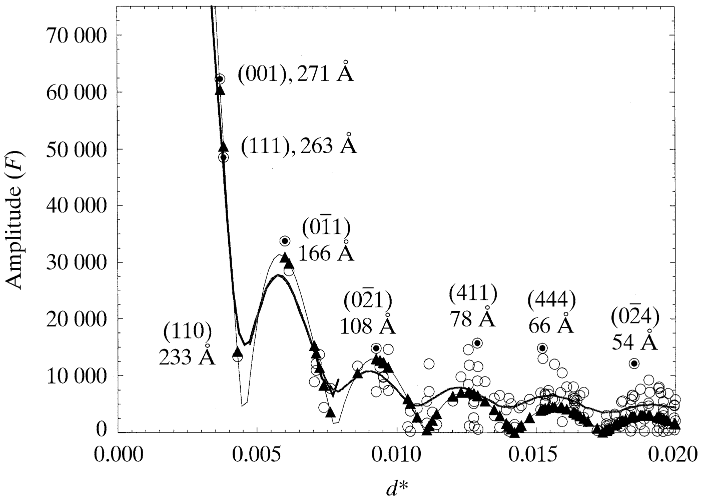


Figure 19.3.2.2

Comparison of the absolute values of single-crystal reflection amplitudes (circles), the solution scattering intensity (thick curve) and the calculated scattering intensity from a uniform sphere (thin curve).

uniform sphere within the particle, and the Fourier transform of such a sphere falls off rapidly if the sphere is relatively large. A typical RNA virus will have an internal RNA core of about 100 Å in radius. The scattering contribution of this region is virtually zero beyond 20 Å resolution. If the low-resolution data are measured accurately, the region occupied by RNA and its general level of interaction with the coat protein can be clearly seen (Fig. 19.3.2.1).

The relation between single-crystal and solution X-ray scattering is clearly illustrated in Fig. 19.3.2.2, where the scattering of a single crystal and the solution scattering of a 320 Å-diameter RNA virus are compared. The figure shows the scattering expected for a uniform sphere 160 Å in radius, the observed solution X-ray scattering from this virus and the single-crystal diffraction from the virus. All three diffraction patterns are in close agreement at resolutions below 50 Å, with the only difference being the continuous sampling of the transform in the computed and solution scattering curve and the discrete sampled transform of the crystalline virus. The RNA density in Flock house virus (FHV) shown in Fig. 19.3.2.1 illustrates the information content of the low-resolution data. DNA density functions in polyoma virus were mapped by employing similar methods (Griffith *et al.*, 1992). Low-resolution single-crystal data are also critical for applying the methods of *de novo* phase determination. Geometric solids, such as a sphere, or a low-resolution electron cryo-microscopy (cryoEM) reconstruction may serve as initial phasing models for such a strategy, and the procedure of refining and extending phases is much more robust if the low-resolution data are measured accurately (Tsuruta *et al.*, 1998).

19.3.3. Solution X-ray scattering studies

Protein crystals have high solvent content: typically 50% and as much as 80% in some cases (*e.g.* Johnson & Hollingshead, 1981; Wikoff *et al.*, 1998). Some solvent is well ordered and visible in the crystal structure. Some, referred to as bulk solvent, is dynamic, not specifically associated with parts of the repetitive lattice and can be readily exchanged within the crystal using a flow cell. The high solvent content of macromolecular crystals makes it probable that most of the features of a protein molecule that are important for its function in solution are preserved in the crystallographic structure. Cases are known, however, in which lattice forces or crystallization conditions have masked functional features of the macromolecule or induced functionally irrelevant oligomerization. A straightforward application of solution scattering is the verification of the relevance of the oligomerization state of a macromolecule observed in the crystal. The methods are complementary, because the crystal structure almost invariably leads to the correct secondary and tertiary structures for the monomer, and the SAXS experiment allows the determination of the quaternary structure of the macromolecule in solution.

The solution scattering pattern of the macromolecule can be calculated directly from a set of atomic coordinates. Scattering computed from models of different oligomerization states or from models of the monomer can be readily compared with the observed scattering pattern. In many cases, not only can the oligomer state be determined, but also the point-group symmetry of the oligomer; hence more information is available than through a solution molecular-mass determination by other physical methods. Using the same approach, effects of various ligands on the state of a protein can be studied in solution. Any large-scale structure alteration of a protein, including changes in quaternary structure or folding and unfolding, can be readily studied in solution, either as a static experiment or as a time-resolved investigation.

Recently, the search for crystallization conditions of macromolecules has been refined by the use of dynamic light scattering. Solution X-ray scattering can be even more effective because it is sensitive to molecular aggregates that are smaller than those detected by light scattering, and this broadens the radius of convergence for conditions that will produce good-quality crystals. For example, the interparticle interaction potential of proteins with a mass of less than 100 kDa can be readily measured in the SAXS experiment, and this has been shown to be an important factor in their crystallization (Bonneté *et al.*, 1997). In a particularly successful experiment, a set of small-angle solution X-ray scattering results was used to improve crystallization conditions, and this lead to single crystals that diffracted beyond 1.9 Å (Mourey *et al.*, 1997). A similar approach was used to distinguish between virus coat protein polymerization and crystallization (Petitpas *et al.*, 1998). Studies in which a number of solution conditions must be tested can be done most effectively using synchrotron small-angle scattering facilities, though it is possible to conduct such studies with a conventional X-ray source.

19.3.3.1. *Information content of solution scattering*

Solution scattering deals with samples that are randomly oriented in a solvent. Three-dimensional information is compressed to one dimension, and a one-dimensional self-correlation function (a rotationally averaged Patterson function) of the protein is obtained when the scattering function is properly Fourier transformed. Excellent books describing the information content of small-angle scattering include those by Guinier & Fournet (1955) and Glatter & Kratky (1982), and for modern perspectives see Feigin & Svergun (1987). Concise reviews with a strong focus on structural biology have also appeared, including those by Trewhella (1998) and Koch (1991). These reviews describe the theoretical background, experimental considerations and ways of extracting structural information. None of these take our perspective, which is addressed to the sceptical crystallographer.

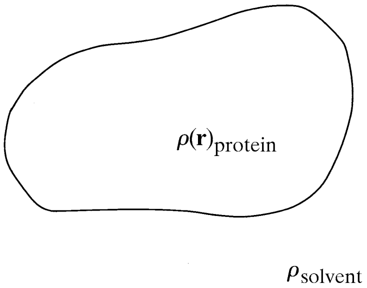


Figure 19.3.3.1
Definition of electron-density contrast. The excess electron density of protein particles over that of the solvent, $\Delta\rho(\mathbf{r}) = \rho(\mathbf{r})_{\text{protein}} - \rho_{\text{solvent}}$, is the origin of solution X-ray scattering.

Assume a monodisperse protein solution. X-ray scattering from the solution originates from the excess electron density of protein particles over that of the solvent, $\Delta\rho(\mathbf{r}) = \rho(\mathbf{r})_{\text{protein}} - \rho_{\text{solvent}}$ (Fig. 19.3.3.1). Note that $\rho(\mathbf{r})_{\text{protein}}$ is typically only about 30% higher than ρ_{solvent}, and that $\Delta\rho(\mathbf{r})$, the electron-density 'contrast', depends on the solvent composition. The scattering intensity observed for such a system is given by

$$\begin{aligned} I(Q) &= F(Q)F^*(Q) \\ &= \int\limits_{\mathbf{r}_1\mathbf{r}_2} \Delta\rho(\mathbf{r}_1)\Delta\rho(\mathbf{r}_2)\exp[-iQ(\mathbf{r}_1 - \mathbf{r}_2)]\,\mathrm{d}V_1\,\mathrm{d}V_2, \end{aligned} \quad (19.3.3.1)$$

where $Q = 4\pi\sin\theta/\lambda$ (2θ is the scattering angle and λ is the wavelength of the X-radiation used). Owing to the rotationally averaged nature of solution scattering, the exponential term may be substituted with the Debye function, $(\sin Q\mathbf{r})/Q\mathbf{r}$, giving

$$I(Q) = \int\limits_{\mathbf{r}} 4\pi\mathbf{r}^2\Delta\rho(\mathbf{r})^2(\sin Q\mathbf{r})/Q\mathbf{r}\ \mathrm{d}\mathbf{r}. \qquad (19.3.3.2)$$

Guinier & Fournet (1955) showed that equation (19.3.3.2) can be approximated by a Gaussian function at relatively small scattering angles resulting in

$$I(Q) = I(Q = 0)\exp(-Q^2R_g^2/3). \qquad (19.3.3.3)$$

This approximation is theoretically valid for $Q \leq 1/R$. Physical parameters of the system that can be determined from the solution scattering experiment are R_g, the radius of gyration of the molecule and $I(Q = 0)$, the scattering intensity at zero angle. The radius of gyration is defined as the root-mean square of the distances of all electrons from their centre of gravity. The radius of gyration can be related to characteristic dimensions of relatively simple objects. For instance, $R_g^2 = (3/5)R^2$ for a sphere of radius R, and $R_g^2 = (a^2 + b^2 + c^2)/5$ for an ellipsoid with semi-axes a, b and c (Feigin & Svergun, 1987). The scattered intensity at zero angle is proportional to the square of the molecular mass of the macromolecule at a defined number concentration. A plot of $\ln I(Q)$ *versus* Q^2 is called a Guinier plot, and if the system is monodisperse and the macromolecule is relatively globular in shape, the plot will be linear up to at least $Q_{\max} = 1/R_g$ (Fig. 19.3.3.2*a* inset). The Guinier plot should begin at a Q value significantly smaller than $1/R_g$. In practice, many investigators take the Q range for the plot to higher Q values such as $Q_{\max} = 1.5/R_g$, or nearly to $2.0/R_g$, since the Guinier plot is often linear to higher Q values. Significant deviation from linearity in the Guinier region is an indication of polydispersity or an extreme deviation from globular shape, such as a cylinder. The solution scattering intensity is proportional to the molecular mass of a particle, thus it is very sensitive to the presence of even a trace amount of large oligomers. Discrete oligomeric states can be identified by the electron pair distance correlation functions as

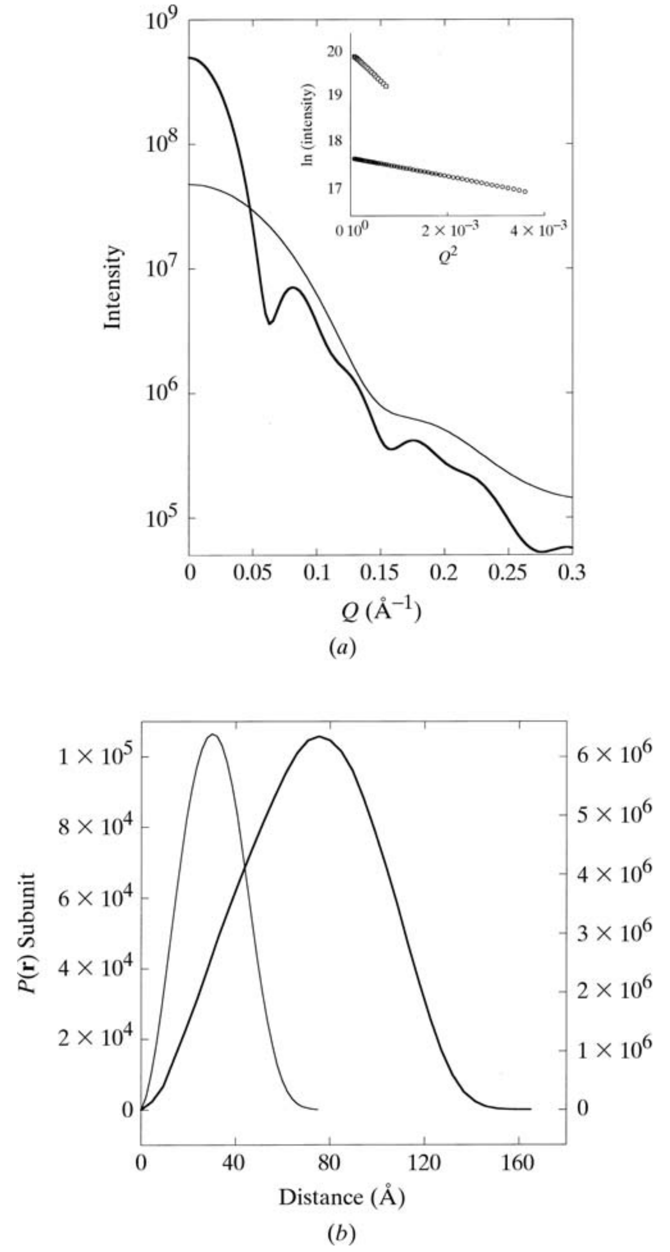


Figure 19.3.3.2
(*a*) The calculated solution X-ray scattering curves from the crystal structure of *Salmonella typhimorium* glutamine synthetase. Glutamine synthetase is composed of 12 identical 469-residue subunits and forms a functional dodecameric assembly (PDB entry 2GLS; Yamashita *et al.*, 1989). The single subunit was computationally isolated, and the solution scattering curves for the subunit (thin line) and the assembled form (thick line) were calculated with a 3 Å-thick hydration layer of 0.03 e Å^{-3} using *CRYSOL* (Svergun *et al.*, 1995). The curves are normalized to an identical weight concentration, that is, the number concentration of the dodecamer is 1/12 of that of the subunit, giving a 12-fold increase in $I(Q = 0)$. The inset shows Guinier plots of the calculated scattering curves. The radii of gyration determined by the Guinier plot to $Q \simeq 1.0/R_g$ (solid-line fit) for the subunit and the assembled form are 24.1 and 55.8 Å, respectively. The 'extended' Guinier plots to $Q \simeq 2.0/R_g$ (circles) give 23.9 and 56.5 Å. Note that both Guinier plots are quite linear beyond $Q \simeq 2.0/R_g$, although the Guinier approximation is theoretically valid to $Q \simeq 1.0/R_g$. It should be noted, however, that the 'extended' Guinier plot tends to give significantly smaller R_g for elongated shapes. (*b*) The electron distance distribution functions. The pair distribution functions $P(\mathbf{r})$ were calculated from the scattering curves in (*a*) using *GNOM* (Svergun, 1992). The maximum intramolecular distance is 75 and 160 Å for the subunit and the assembled enzyme, respectively. The $P(\mathbf{r})$ gives the radius of gyration as 23.6 Å for the subunit and 55.0 Å for the assembled enzyme.

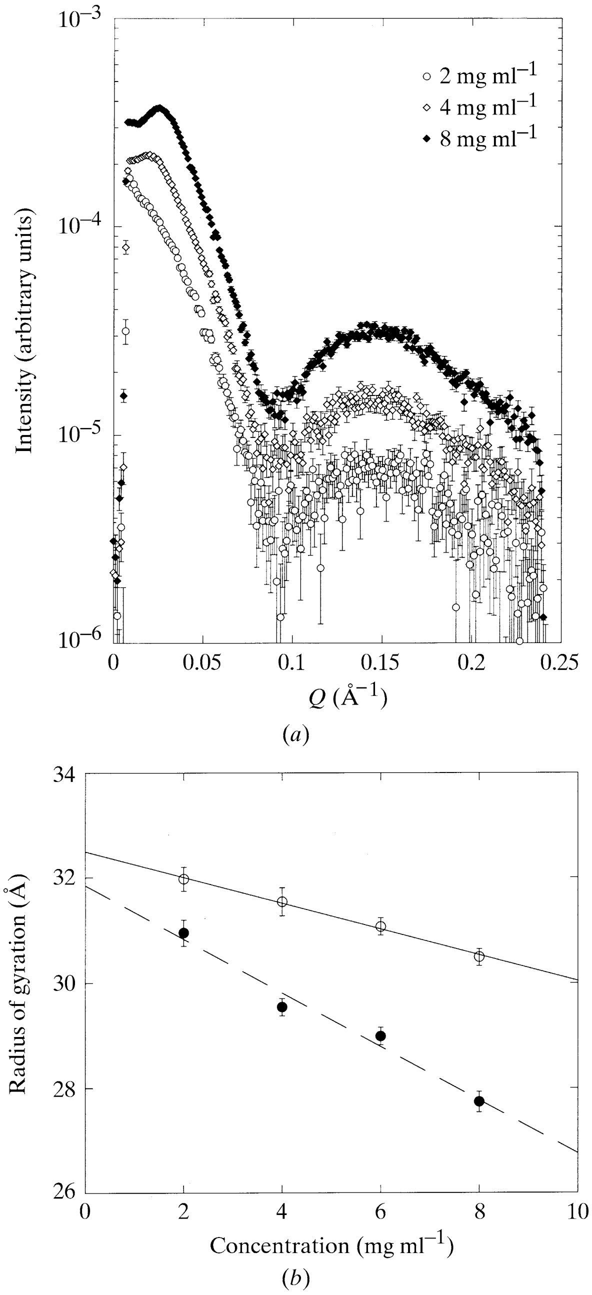


Figure 19.3.3.3
(*a*) Interparticle interference in biological small-angle scattering. Scattering from myelin P0 glycoprotein solubilized in 0.1% SDS, recorded at 2, 4 and 8 mg ml^{-1} (Inouye *et al.*, 1999). (*b*) Plots of measured radii of gyration of *E. coli* aspartate transcarbamoylase catalytic trimer *versus* concentration in the absence (open circles) and presence (filled circles) of the bisubstrate analogue *N*-phosphoneacetyl-L-aspartate (E. R. Johnson, H. Tsuruta & H. K. Schachman, unpublished data).

shown in Fig. 19.3.3.2(*b*). A second contribution to nonlinear Guinier plots is interparticle interference due to strong interactions among particles. In these cases, the particle correlation function is overlaid on the scattering function from the molecular envelope. Strong interparticle interference is illustrated in the example in Fig. 19.3.3.3. The P0 glycoprotein was solubilized with sodium dodecyl sulfate (SDS), whose net charge expels solubilized protein particles from each other, generating a correlated interparticle distance in the range of 250 Å. While the correlation is rather weak at 2 mg ml^{-1}, much stronger interference is seen at higher concentrations. This is an extreme example of interparticle interference, because a noticeable interference peak is observed in a much higher concentration range, well over 10 mg ml^{-1}, for most soluble proteins. Smaller proteins often exhibit more pronounced interparticle interference at lower concentrations than larger proteins, perhaps because the number concentration of protein molecules is higher with smaller proteins. In the case of the 100 kDa protein in Fig. 19.3.3.3(*b*), no detectable interference peak is seen. While Guinier plots are linear throughout the concentration range studied, the slope, *i.e.*, the radius of gyration, gradually decreases as the protein concentration is increased (data not shown). Also note that a small inorganic molecule could change the interparticle interference, as seen in slight change in the slope of the R_g *versus* concentration plot. This small ligand induces a conformational change in the protein, which alters its electrostatic properties. The best results are usually obtained when scattering curves are recorded at a few different protein concentrations and then extrapolated to infinite dilution. This procedure generates a scattering curve free from interparticle interference. Extrapolated values of R_g are often used to assess the magnitude of protein conformational changes. The interparticle interference, on the other hand, could be used to evaluate electrostatic properties of the protein, though this effect has not yet been clearly linked to any specific physical parameter, except the second virial constant. Note that the positive slope of the R_g *versus* concentration plot is an indication of either concentration-dependent oligomerization or non-specific aggregation. Different solution conditions should be tested to avoid a positive slope. The forward scattering intensity $I(0)$ is calculated by extrapolating the Guinier plot to $Q = 0$. $I(0)$ is proportional to the number of electrons in a particle, though it depends on the solvent electron density, given the definition of electron-density contrast. This complicates the determination of absolute molecular weight. Relating the measured $I(0)$ values to the molecular mass of the protein thus requires a well characterized monodisperse protein system with a known molecular weight for calibration. A solution of hen egg lysozyme or bovine serum albumin is often used. These proteins, however, tend to show significant changes in scattering intensity at very small angles due to radiation-induced aggregation under a full synchrotron-radiation beam. Synthetic polymers such as polyethylene are convenient standards because of their high stability; however, they scatter X-rays orders of magnitude more than many protein solutions, making it difficult to scale the data to such a standard. The problem is not usually serious, as most biological studies deal with proteins of known molecular mass. Studies of oligomeric proteins are simplified, because a species of known molecular mass, *e.g.*, a monomer, usually serves as an internal standard. The Guinier plot $\ln I(Q)$ *versus* Q^2 is no longer linear, even at small angles, when the shape of the protein deviates significantly from a globular shape. A plot of $\ln QI(Q)$ *versus* Q^2 gives a radius of gyration of the cross section and mass per unit length for a rod-like particle as well as a radius of gyration of thickness and mass per unit area (Feigin & Svergun, 1987; Glatter & Kratky, 1982).

Other information obtained directly from the experiment when there is no *a priori* knowledge of the structure is the electron pair distribution function $P(\mathbf{r})$, *i.e.*, the histogram of distances of all possible pairs of electrons within a particle (Fig. 19.3.3.4). This quantity is obtained from the Fourier transform of the observed scattering intensities. Fourier termination effects and smearing of scattering curves due to the use of a nonideal experimental geometry must be considered in the derived result. Algorithms have been developed to incorporate the experimental geometry in the back Fourier transform (Glatter, 2004; Moore, 1980). The use of nearly parallel, focused synchrotron X-ray beams, with a small beam cross section and a long sample-

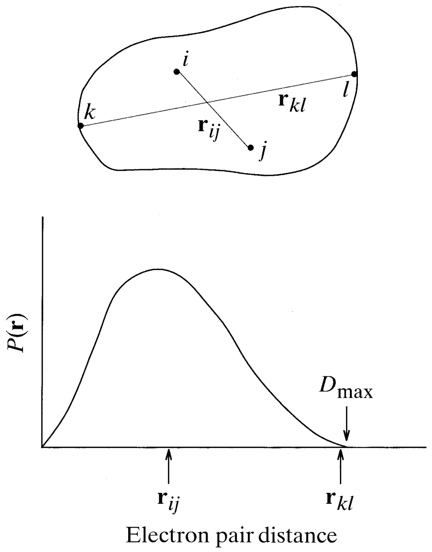


Figure 19.3.3.4
Definition of the electron pair distance distribution function $P(\mathbf{r})$. It is a histogram of all electron pair distances within a particle as depicted here. The largest dimension of the molecule $D_{\max}$ is determined from the point where $P(\mathbf{r})$ falls to zero.

to-detector distance, often eliminates the need for desmearing scattering curves and this leads to $P(\mathbf{r})$ functions with minimal errors from data correction and manipulation. $P(\mathbf{r})$, in turn, gives R_g and $I(Q = 0)$:

$$R_g = (1/2) \int^{D_{\max}} \mathbf{r}^2 P(\mathbf{r})\, \mathrm{d}\mathbf{r} \Big/ \left(\int^{D_{\max}} P(\mathbf{r})\, \mathrm{d}\mathbf{r} \right), \qquad (19.3.3.4)$$

$$I(0) = 4\pi \int^{D_{\max}} P(\mathbf{r})\, \mathrm{d}\mathbf{r}. \qquad (19.3.3.5)$$

The entire scattering curve is used to obtain $P(\mathbf{r})$, thus giving more accurate values for R_g and $I(0)$ than those from a Guinier plot, even when the data do not cover many of the scattering angles below $Q = 1/R_g$. The values of R_g obtained from $P(\mathbf{r})$ tend to be less sensitive to trace amounts of contaminating higher aggregates.

19.3.3.1.1. Solution scattering and crystal structures

SAXS is an extremely powerful method for studying the relative change in molecular weight, and thus the oligomerization state, of a protein system. Such studies are useful in the case of proteins that change in quaternary structure as a means of regulating biological function. Changes in association state such as $n\mathrm{A} \leftrightarrow \mathrm{A}_n$ lead to a large change in $I(0)$, which is proportional to the square of the molecular mass and to the number of particles in solution. In the example illustrated in Fig. 19.3.3.2, the formation of a dodecameric assembly composed of 12 identical subunits at a constant weight concentration results in the increase in $I(0)$ by a factor of 12.

The radius of gyration can be readily calculated with atomic coordinates, employing a program such as *X-PLOR*. This provides a quick means of verifying the overall crystal structure in solution. R_g values measured in solution are often larger, by up to a few ångstroms, than those calculated from atomic coordinates. Such differences are larger than statistical errors in a carefully planned set of experiments, although trace amounts of molecular aggregates due to the presence of naturally present higher oligomers, impurities or radiation-induced aggregation could increase apparent radii of gyration. The apparent protein envelope may be enlarged significantly by the presence of structured water molecules surrounding the protein. Strongly associated water molecules on the protein surface are significant, because the electron-density contrast between a protein and a solvent is small (Lattman, 1989). A recent study by Svergun, Richards *et al.* (1998), combining X-ray and neutron solution scattering techniques, verified that the hydration layer is critical when interpreting solution X-ray scattering curves based on crystal structures, because crystallographic structures underestimate the extent of the hydration layer. This layer is about 3 Å thick, with an average density about 10% higher than that of bulk solvent. The program *CRYSOL*, developed by Svergun *et al.* (1995), has incorporated the first hydration layer in the calculation of scattering curves and radii of gyration, providing estimates of the scattering close to those observed experimentally.

The Debye formula, (19.3.3.6) below, demonstrates that solution scattering intensity is sensitive to the overall shape of the molecule and can readily follow changes in particle dimensions. The intensity of scattering is given by

$$I(Q) = \sum_i \sum_j \Delta\rho(\mathbf{r}_i)\Delta\rho(\mathbf{r}_j) \sin[Q(\mathbf{r}_i - \mathbf{r}_j)]/Q(\mathbf{r}_i - \mathbf{r}_j), \quad (19.3.3.6)$$

where the indices represent individual mass elements or electrons. The solution scattering intensity from a macromolecule of known structure can be calculated readily with this expression. An entire scattering curve can be computed rapidly, even for molecules with thousands of atoms. For instance, calculation of the scattering curve of a glutamine synthetase subunit with 3636 atomic positions took approximately 30 min with a DEC Alpha workstation. The calculation of the dodecameric assembly of glutamine synthethase was performed using only α-carbon positions, instead of all atomic positions, to reduce computation time. This approach still approximates the molecular envelope of this large molecular complex very well, and the effect of the hydration layer could be incorporated, as proposed by Lattman (1989). Alternatively, three-dimensional objects can be approximated by a sum of spherical harmonics (Harrison, 1969; Stuhrmann, 1970), and the Fourier transform of the spherical harmonic expansion is an effective method for evaluating the Debye formula. The program *CRYSOL*, developed by Svergun *et al.* (1995), can compute the entire scattering curve for the dodecameric glutamine synthetase in 10 min using this method. Explicit incorporation of symmetry allows a solution scattering curve from giant molecular complexes, such as an icosahedral virus particle, to be computed efficiently (Zheng *et al.*, 1995).

19.3.3.1.2. Low-resolution model determination using solution scattering

The coordinates for a structure are often unavailable, but its general shape is known from electron microscopy or hydrodynamic studies. Pilz *et al.* (1979) demonstrated that reasonable scattering patterns from complex systems could be generated using a model that approximated the shape of a large complex with a number of uniform-density spheres. This approach has been improved over the years to make it less dependent on initial models. For instance, a genetic algorithm was recently used to select a few best models composed of uniform-density spheres to generate density functions in close agreement with solution scattering data (Chacón *et al.*, 1998). A program for this approach, *DALAI_GA*, has recently been released. There have been other effective approaches to the determination of low-resolution structures from solution scattering, including Monte Carlo methods (Olah *et al.*, 1995) to select the most plausible model among many possibilities. On the other hand, the program

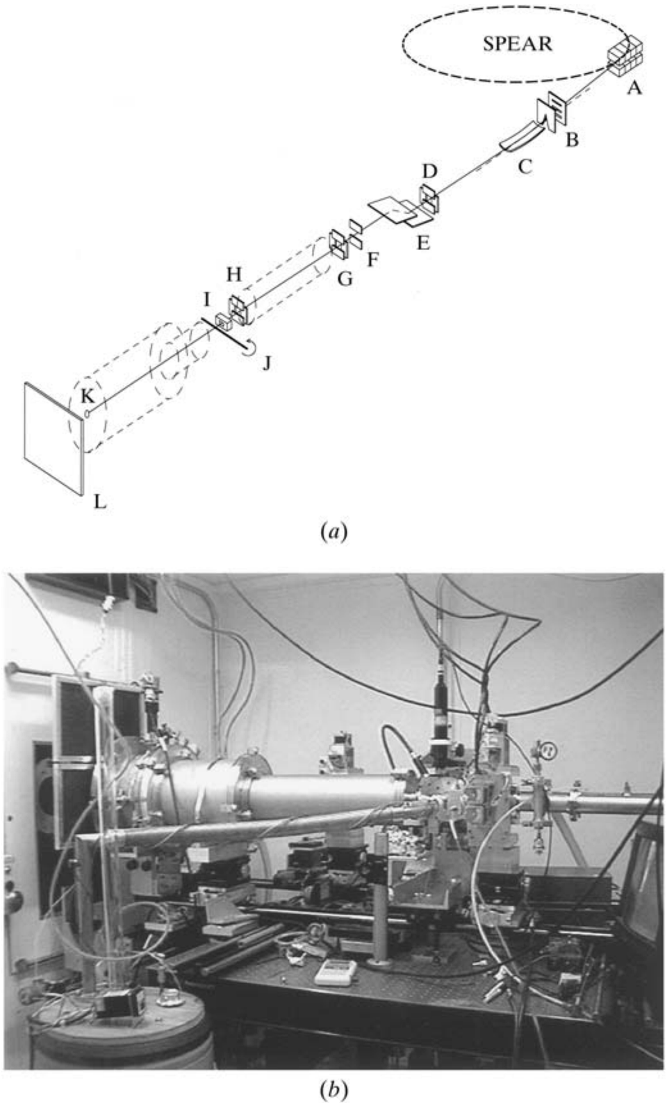


Figure 19.3.3.5
A small-angle X-ray scattering instrument on BL 4-2 at the Stanford Synchrotron Radiation Laboratory. (*a*) A diagram of the instrument composed of an eight-pole wiggler source (A), mirror slits (B), toroidal focusing mirror (C), monochromator slit (D), double-crystal monochromator (E), fast beam shutter (F), beam-defining slit (G), guard slit (H), ion chamber (I), crystal spindle axis (J), which would be replaced with a solution sample cell for solution X-ray scattering, beam stop (K) and an image plate detector (L), which would be replaced with a gas-chamber detector for solution scattering. Vacuum beam flight paths are drawn in dotted lines. (*b*) A view of the instrument as configured for small-angle single-crystal diffraction with a crystal-to-detector distance of 1.3 m. Some of the components in (*a*) are also seen in (*b*).

DAMMIN uses simulated annealing for optimization (Svergun, 1999).

Multiple-sphere modelling approaches involve many parameters, and the verification of resulting models can be difficult. This is a fundamental problem in solution scattering, because random orientation of a protein in solution limits the number of structural parameters that can be determined reliably. The scattering curve recorded in the resolution range 700 to 20 Å, a typical angular range covered in solution X-ray scattering studies of proteins, allows only a small number of parameters to be refined, typically less than 20, according to Shannon's information theorem. Density functions for oligomeric proteins displaying specific molecular symmetry can be computed directly from the solution scattering pattern with spherical harmonic expansions that reflect the symmetry (Jack & Harrison, 1975). A set of solution scattering curves from components of an icosahedral virus particle (Schmidt *et al.*, 1983) was analysed by an icosahedral harmonic expansion technique, directly generating a low-resolution structure of the virus particle that was comparable to one determined by cryo-electron microscopy and image reconstruction (Zheng *et al.*, 1995). More recently, density functions for asymmetric particles were derived *de novo* from the solution scattering patterns with a spherical harmonic expansion and no imposed symmetry (Svergun *et al.*, 1996). An interactive modelling tool has been developed that combines the spherical harmonics approach with placement of protein-subunit coordinates in real space to construct a model for a large macromolecular complex using solution scattering data (Kozin *et al.*, 1997).

19.3.3.2. *Instrumentation for small-angle X-ray scattering*

A small-angle X-ray scattering instrument typically consists of an X-ray source, a set of X-ray optics, a sample holder and a detector system (Fig. 19.3.3.5). The source can be an X-ray tube, a rotating-anode source or synchrotron radiation. The choice of optical components depends on the type of source, beam-flux requirement and angular range to be covered. The optics system includes a beam focusing device, such as a mirror, a monochromator, beam-collimation slits, vacuum paths and a beam stop. Owing to the very weak level of X-ray scattering from solutions of biological macromolecules, caution must be taken to minimize scattering from air and window materials used in the sample holder as well as to contain elements of the system under vacuum. The choice of detector depends primarily on the level of signal expected for typical samples at the detection plane. Other factors that are crucial in choosing a detector system include active area, spatial resolution and read-out speed. The level of noise generated within an entire detector system must always be constant and preferably much lower than the weakest level of signal to be measured.

19.3.3.2.1. *Instruments on conventional sources*

Kratky cameras, which are commercially available, have been used for many years in small-angle X-ray scattering studies of synthetic polymers and relatively large biological systems (Glatter & Kratky, 1982). These instruments record scattering in only one dimension, thus they are not always suitable for the study of weak X-ray scatterers, although excellent accessibility to small angles is often achieved. More recent small-angle X-ray scattering instruments have a pinhole collimation system similar to those used on synchrotron instruments described below. They allow isotropic scattering to be measured with a two-dimensional detector (Bu *et al.*, 1998). Synthetic multilayered materials, such as Mo-B_4C, formed on a figured surface serve as a monochromator element as well as a focusing device and produce an X-ray beam with very small divergence (Schuster & Göbel, 1995). Many instruments on conventional sources could benefit from this new development in X-ray optics.

19.3.3.2.2. *Synchrotron instruments*

The needs for time-resolved capability and for measuring weak X-ray scattering have lead to the development of synchrotron small-angle X-ray scattering instruments. Most of these instruments employ the pinhole camera geometry and are suitable for fibre diffraction experiments as well. [For a review, see Koch (1988).] Major synchrotron facilities that produce radiation in the hard X-ray regime have at least one small-angle X-ray scattering instrument of this type. Gas proportional counters are common with these instruments, as they allow photon counting virtually without added noise (Petrascu *et al.*, 1998). Gas proportional counters are equipped with very fast electronics modules, allowing more than 1000 time-resolved scattering patterns to be recorded successively every second (Boulin *et al.*, 1988). The disadvantage of these detectors is the count-rate limit, *i.e.*, counting efficiency drops to below 50% at about 10^5 to 10^6 counts per second. This is due to the fact that the position of photon arrival is converted to a time domain, which is then registered in a histogram memory module. The space-to-time conversion occurs first on the delay lines incorporated in the gas-chamber detector, then the time-to-space conversion is performed by time-to-digital converters. These processes take a fraction of a µs per X-ray photon using present technologies. The position of the X-ray photon cannot be recorded effectively when another photon is being processed, thus leading to the count-rate problem mentioned above. Cipriani *et al.* (1994) proposed a new delay-line technology to increase this limit by a factor of at least 10, making gas-chamber detectors useful at high count rates. Experimenters should, however, be aware of space-charge effects in gas proportional counters. Two experimental artifacts can be observed. First, intensity near the beam stop is reduced when macromolecular complexes are studied. Second, peaks with high intensity can be reduced at their maximum, giving the impression of two closely spaced weaker peaks with a trough at the location of the true peak. Arriving photons are not counted by the detector when the local count rate exceeds the recombination rate of a detector gas molecule. Modern integration detectors, such as the image plate and the charge-coupled-device (CCD)-based detectors, have virtually no count-rate limit and have been characterized for small-angle scattering. These new detectors have certain limitations, such as a relatively slow data-acquisition rate, a problem for time-resolved studies. In general, care should be taken to match the detector to the experiment.

19.3.3.3. Experimental considerations

19.3.3.3.1. Sample preparation

Sample volumes required for one measurement are between 10 to 50 µl, occasionally more, depending on the specific design of the sample cell used. The concentration required to record a scattering curve with satisfactory statistics depends primarily on the molecular weight of the protein and the beam flux. Approximately 1 mg ml^{-1} of a small protein (10–20 kDa) is the lower limit for recording a scattering pattern with satisfactory statistics when experiments are performed with a typical synchrotron instrument equipped with a gas-chamber detector. Somewhat lower concentrations of larger molecular weight proteins may be used. Higher concentrations will improve statistics significantly and reduce exposure times, but interparticle interference may result from high concentrations. Time-resolved experiments benefit dramatically from higher sample concentrations. In addition to the scattering power of the sample, the signal-to-background ratio and overall stability of an instrument (from X-ray source and optics to detector) limit the lowest concentration for a given experiment. Although higher concentrations add dramatically to the scattering and improve statistics, sample solutions must be monodisperse. Small-angle solution scattering is not well suited to the study of polydisperse systems, which give scattering of the entire molecular population weighted by the square of the mass, although a few distinct populations of substantially different sizes may be resolved with good-quality data. Chemical components that may have been carried along in a sample preparation, such as ammonium sulfate, sucrose, chloroform or caesium chloride, should be removed. The presence of such compounds may change the electron-density contrast and X-ray absorbency of the sample. In general, this can be most effectively avoided by exhaustive dialysis with the desired buffer solutions. The outer solution used for the final dialysis should be used for the blank measurements. Scattering contributions from the buffer solution, the sample cell and parasitic scattering must be subtracted from the measured scattering curve; these can be measured accurately from a well prepared blank. Extra buffer solution should be available for sample dilution. The data quality is improved and problems with radiation-sensitive samples are readily detected when protein concentrations and biological activities of samples are measured before and after the scattering experiment. Accurate protein concentration measurements permit scattering intensities from different samples to be scaled together accurately. This is particularly important in determining molecular weight.

19.3.3.3.2. Sample-handling devices

Sample holders used in solution scattering are either Lindeman glass or quartz capillaries, or machined cells equipped with flat windows (Fig. 19.3.3.6). Glass capillaries have been widely used to contain a sample solution. A beam size significantly smaller than the diameter of the capillary is required to minimize strong parasitic scattering from the round edge of the capillary. A large beam size in the horizontal direction may be used with capillaries to obtain stronger scattering intensity without adding parasitic background. Capillary cells are suitable for measuring scattering primarily in the direction perpendicular to the long axis of the cell. The advantage of the capillary cell is the small volume of sample solution required for measurements. About 4 µl of a sample solution in a 1 mm-diameter capillary provides sufficient material for the experiment. By employing a specialized holder, a capillary cell can be placed under a vacuum, minimizing air scattering (Dubuisson *et al.*, 1997). This feature is useful for solution scattering studies of small proteins. Anaerobic samples may be sealed in a capillary.

Another common sample cell holds the solution between a pair of flat windows. This cell offers two improvements over capillary cells – a larger beam cross section may be used, which increases the number of photons incident on the sample, and a two-dimensional detector can be effectively used for recording the scattering. Flat-window cells are available that require only about 10 µl of solution. Both capillary and flat-window cells require a holder with temperature regulation. The choice of window materials for the sample container is important because of the weak sample scattering. An X-ray-transparent material is required that has little intrinsic scattering within the scattering-angle range of the sample. Common window materials include synthetic mica of high purity and certain types of polypropylene and polyamide. Etched high-quality beryllium has also been used. The windows must be thin enough to transmit X-ray photons, but rigid enough to keep a constant beam-path length through sample solutions. A valuable comparison of window materials has been published recently (Henderson, 1995). A protein solution scattering pattern and a corresponding blank scattering curve should be measured in the same sample cell, unless the cells and windows are identical.

Special sample-handling devices have been built for time-resolved studies. Stopped-flow mixers for solution X-ray scattering are routinely used by a number of research groups (Kihara,

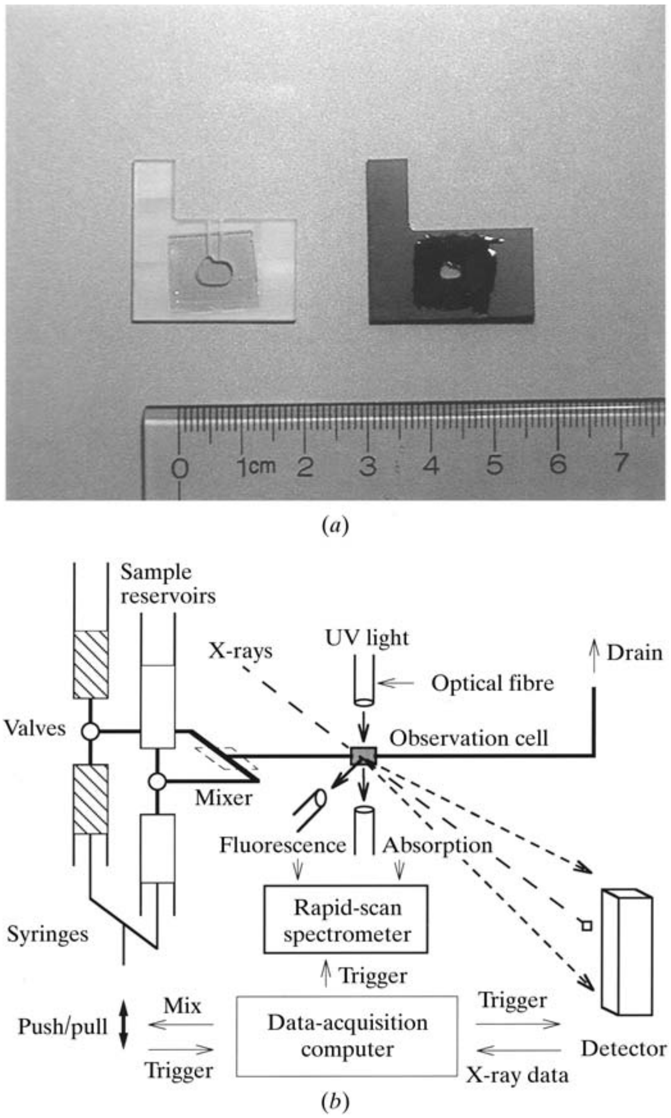


Figure 19.3.3.6

(*a*) Flat-window cells for solution scattering and (*b*) a diagram of a stopped-flow rapid mixer for time-resolved solution scattering. The solution cell to the left in (*a*) is made of polycarbonate and is equipped with two synthetic mica windows. A sample solution is injected through one of two sample loading channels using a microsyringe. The black cell to the right is made of coloured polyoxymethylene for light-activated, time-resolved studies and has a smaller sample chamber. In (*b*), two solutions, *e.g.* an enzyme solution and a substrate solution, are put in sample reservoirs, loaded into individual syringes and wait for a trigger signal from the data-acquisition system. Then the two solutions are rapidly mixed, transferred to the observation cell, typically within 5 ms or less, and a trigger signal initiates a series of time-sliced scattering-data acquisitions. This stopped-flow mixer is also equipped with optical paths to monitor absorption or fluorescence from the protein solution in the observation cell.

1994), and a high-pressure solution cell has been constructed (Czeslik *et al.*, 1996). The assembly process of tobacco mosaic virus has been studied using a temperature-jump apparatus (Hiragi *et al.*, 1988).

19.3.3.3.3. *Designing experiments*

The angular range for data collection should first be determined on the basis of the structural information required, and an instrument should be configured accordingly. Most camera systems have the flexibility to use more than one sample-to-detector distance, and they are occasionally equipped with an additional detector to cover higher angles.

The exposure time is usually determined empirically, taking into account the data statistics required, the number of conditions to be investigated and the total amount of beam time available. Typical exposure times run from a few minutes with a synchrotron source to several hours with a laboratory source. Exposure time depends strongly on the sample molecular weight, the concentration and the angular scattering range of interest. It is essential to measure the blank solvent, *i.e.*, the buffer solution in which the protein of interest is dissolved. It should be recorded with the same statistical significance as the protein solution scattering. This blank is subtracted from the observed protein solution scattering intensity, and therefore its intensity contributes to the overall counting statistics. Accurate blank measurements can take significant amounts of time and should not be ignored when planning the experiment. Blank solution scattering patterns should be recorded as often as possible. They serve as an internal control to detect systematic errors during periods of instrumental instability. In addition, the intensity of the incident X-ray beam should be integrated during the exposure time and used to scale the scattering data. Although it is useful to record the beam intensity transmitted through the solution sample, absorption is generally not a significant problem for solution scattering.

19.3.3.3.4. *Data-collection practices*

Scattering intensity measurements should be performed with a series of experiments. Solutions at different concentrations will demonstrate that the anticipated trends are observed at a qualitative level, even before the data are processed. Time-dependent aggregation or degradation are detected by recording scattering curves in a series, with a short exposure for each curve (*e.g.*, recording 10 successive scattering curves every 30 s for one protein solution). Proteins in solution may degrade in the full flux of a synchrotron-radiation beam and may exhibit radiation-induced aggregation. This is easily recognized by inspecting the very small angle region of scattering patterns that are recorded successively using short exposure times. Radiation damage can be reduced or eliminated either by attenuating the primary beam or by adding small amounts of antioxidants, which may help remove free radicals. There are several cases in which a small amount of dithiothreitol or β-mercaptoethanol prevented radiation-induced aggregation. The effects of the sulfur-reducing agents are consistent with the observation by a time-resolved single-crystal study that the sulfur atom disappears first from an electron-density map (Weik *et al.*, 2000). Temperature control is essential for most experiments. Irradiation by a 9 keV X-ray beam at a rate of 10^{11} photons s^{-1} will heat up an aqueous solution of a few µl by a fraction of a Kelvin per second if a constant-temperature device is not employed.

Preliminary data processing should take place immediately with each sample measurement. Processing 'on the fly' is particularly important when experiments are performed at a synchrotron where access may be limited. Visual inspections or statistical evaluation should be made for time-dependent aggregation at a low scattering angle. A determination of the sample molecular weight by estimating $I(0)$ and the radius of gyration from the Guinier plot would also be useful. In addition to making measurements with a blank, scattering from a standard sample should be measured, allowing the relative molecular weight of the unknown to be determined. A standard, stable, well defined known sample should be recorded prior to any other measurements. Running the same standard sample in every data-collec-

Table 19.3.3.1
List of commonly used software for solution scattering

Program	Function	Reference
OTOKO†	Data evaluation for noncrystalline diffraction	Koch (1988)
SAPOKO	Data evaluation for solution scattering	Koch & Svergun (1992)
GNOM	Indirect Fourier transform for $P(\mathbf{r})$	Svergun (1992)
IFT	Indirect Fourier transform for $P(\mathbf{r})$	Glatter (2004)
CRYSOL	Calculation of $I(Q)$ from PDB files	Svergun *et al.* (1995)
SASHA	Spherical harmonics structure determination	Svergun, Volkov *et al.* (1997)
ASSA and *ALM22INT*	Three-dimensional modelling in real space	Kozin *et al.* (1997)
DAMMIN	Simulated annealing structure determination	Svergun (1999)
DALAI_GA	Genetic algorithm structure determination	Chacón *et al.* (1998)

† An X-terminal compatible version of *OTOKO* has been developed at Daresbury Laboratory.

tion period makes it easier to compare and scale data sets recorded at different times. Sample solutions should be recovered from sample cells and stored separately, so that post-irradiation concentrations can be measured, as well as their biological activity if this can be assayed. A standard sample that gives sharp diffraction peaks of known spacing should be measured to allow the conversion of the detector channel number to Q values. A dried collagen fibre or a non-biological or more stable specimen such as cholesterol myristate is commonly used. Determining the direct beam position at the detector plane is critical in converting a detector pixel number to a Q value. The symmetric centre of a diffraction pattern of a powder or fibre sample may be used, or a thick metal foil may be inserted to attenuate the primary beam intensity and the beam position measured directly with the detector. In order to place the scattering intensity on an absolute scale, *i.e.*, to determine the scattering cross section, data from a calibrated standard sample need to be recorded.

19.3.3.3.5. *Data processing and analysis*

Processing the solution scattering data first requires scaling the raw intensity measurements to the incident or transmitted beam flux, then averaging scattering curves recorded from identical samples and subtracting measurements from the blanks described above. Corrections for detector non-uniformity or image distortion, if any, must be made, and then the detector channel numbers are converted to Q values (or inverse Bragg-scattering spacing) through the use of the scattering standard for absolute scattering-angle determination. Many facilities provide data-processing support at least to this level. Except for the image-distortion correction, commercially available analysis and visualization software can be used for this purpose. Data formats currently vary among facilities and depend on the detector systems used, but efforts are underway to establish a standard format that would enable the wide use of common data-reduction software. Common formats will make it straightforward to perform data reduction 'on the fly' and to monitor the quality of the data closely. Unfortunately, the size of the biological small-angle scattering community is not conducive to costly development of commercial software, so user friendliness has not generally been a priority in its development.

Plots of I *versus* Q that are corrected for all the known experimental factors are obtained from the data processing, and structural parameters may be derived from these data. The molecular weight and the radius of gyration can be derived immediately from the Guinier approximation. Calculation of the optimal electron pair distribution function by computing the Fourier transform of the intensity function provides the only 'real-space' data directly obtainable from the experiment. The program *GNOM* developed by D. Svergun is widely used for this purpose and may be obtained from the program author. An algorithm developed by P. Moore has been widely implemented, and software developed by Glatter *et al.* (Glatter, 2004) is available. Table 19.3.3.1 lists software suites frequently used in interpreting solution X-ray scattering data.

19.3.3.4. *Recent applications of solution X-ray scattering in structural molecular biology*

Solution X-ray scattering provides a direct means of measuring shape at low resolution, *i.e.*, the radius of gyration, the maximum dimension within a protein and the relative molecular weight. Determining the oligomeric state of a protein system is straightforward. Dimensional parameters for simple objects can be obtained indirectly by fitting the observed pattern with spherical or higher-symmetry harmonics. The concept can be extended by use of partial structural models obtained by crystallography and NMR to construct a low-resolution model of large macromolecular complexes that are not amenable to these high-resolution techniques. The refinement of such models against the solution scattering data of the putative complex bridges between atomic resolution structures and biological functions performed by large complexes. Time-resolved studies expand knowledge derived by a variety of structural techniques into the context of experimental molecular dynamics. The examples given below demonstrate some such complementing aspects of solution X-ray scattering studies.

19.3.3.4.1. *Studies of proteins in solution that complement high-resolution structures*

Recent advances in solution scattering data interpretation have made it possible to determine three-dimensional structures at resolutions comparable to those of electron microscopy. In some cases, low-resolution structures were derived from solution X-ray scattering independent of a structural model obtained by other structural techniques. Svergun *et al.* (1994) derived a 15 Å-resolution model of 50S ribosome using the spherical harmonics method. The low-resolution structures of the dimeric and tetrameric forms of pyruvate decarboxylase were obtained by the same method from X-ray solution scattering data (König *et al.*, 1992) before the crystal structure was published. The X-ray structure, at a comparable resolution, was in close agreement with the solution scattering result. The same enzyme obtained from different organisms has also been modelled with solution scattering data.

Calcium-binding proteins have been extensively studied by solution X-ray scattering. These data first demonstrated the large conformation change of calmodulin promoted by binding Ca^{2+}

(Seaton *et al.*, 1985). Comparison of the solution scattering curve calculated for the crystal structure of calmodulin with the experimental scattering curve showed that crystal-packing forces substantially altered the solution conformation of the protein (Heidorn & Trewhella, 1988). Kataoka *et al.* (1989) reported the first glimpse of the conformational change of calmodulin induced by melittin, a model peptide for target enzymes. These early studies led to a number of important solution scattering studies on the protein–protein interaction in Ca^{2+}-regulated molecular switching, including the troponin system in muscle contraction. Krueger *et al.* (1997) recently took a combined approach to X-ray and neutron small-angle scattering by using contrast variation to obtain the first structural model of calmodulin complexed with an enzymatically active truncation mutant of skeletal muscle myosin light chain kinase.

A series of structural studies on 70 kDa heat-shock cognate protein, a molecular chaperone, combined crystallography and X-ray solution scattering. It was not possible to crystallize the whole protein, but the crystal structure of the ATPase domain was solved. ATP binding induces a conformational change in the protein, resulting in the release of a bound peptide. Wilbanks *et al.* (1995) constructed a low-resolution model of the whole molecule from solution X-ray scattering data, based on the ATPase domain crystal structure. Solution X-ray scattering was recently used to screen a point-mutated version of the protein, which retains the ATP-induced conformational change, contributing to the interpretation of the role of specific residues in the molecular chaperone mechanism (Sousa & McKay, 1998).

Solution scattering complemented high-resolution NMR structural studies in the investigation of titin (connectin), a giant muscle protein (Improta *et al.*, 1998). The high-resolution structures were determined for two immunoglobulin-like fragments of the I-band region of titin. Two ellipsoids that simulate the molecular envelopes of the two fragments were then used to model solution X-ray scattering data in order to determine the relative position of the two fragments. The resulting structural model suggests that the motions around the interdomain-connecting regions are restricted and that titin behaves as a row of beads connected by rigid hinges. A similar approach was taken for EGF domains in coagulation factor X (Sunnerhagen *et al.*, 1996).

There is a long list of large proteins or protein complexes whose solution X-ray scattering studies have made important contributions in structural biology. One study demonstrated that the quaternary structure of an allosteric enzyme, *E. coli* aspartate transcarbamoylase, in the R (relaxed) state is significantly different from that observed in the R-state crystal, suggesting that a crystal lattice could deform functional quaternary structure (Svergun, Barberato *et al.*, 1997). *A. vinelandii* nitrogenase, the key enzyme in nitrogen fixation, is composed of two proteins: Fe protein and MoFe protein, which were crystallized separately. Grossman *et al.* (1997) determined a low-resolution structure of the complex of Fe and MoFe proteins stabilized by ADP–AlF^{4-}, a nucleotide triphosphate analogue, solely from solution X-ray scattering data. This was later confirmed by a crystallographic study of the same complex (Schindelin *et al.*, 1997). A more advanced approach was used to construct the three-dimensional structure of *E. coli* F_1 ATPase, which consists of four different soluble subunits in the form of $\alpha_3\beta_3\varepsilon\delta$ (Svergun, Konrad *et al.*, 1998). Two smaller subunits, ε and δ, were disordered in the crystal structure, while NMR provided solution structures for these subunits. Solution scattering was used to construct the three-dimensional structure of the F_1 ATPase structure that incorporates all the subunits. The program *CRYSOL* was used to calculate individual scattering amplitudes by approximating high-resolution structures with sums of spherical harmonics. Then individual amplitudes were combined, according to the relative positions of all subunits, while the subunits were moved with respect to each other to fit the experimental scattering amplitudes obtained for the whole complex. The programs *ASSA* and *ALM22INT* were used to construct the most plausible model for the complex. A 32 Å-resolution structure of *M. Sextra* V_1 ATPase was determined *ab initio* using the spherical harmonics method (Svergun, Aldag *et al.*, 1998). Here, a three-fold molecular symmetry in the A_3B_3 part, which resembles $\alpha_3\beta_3$ in F_1 ATPase, reduced the number of parameters. The model structure of V_1 ATPase has a 110 Å-long stalk region, corresponding to the total mass of the $CDEFG_3$ part of the complex that connects V_1 to the V_0 domain embedded in the membrane.

X-ray scattering amplitudes have recently been used in the correction of the contrast-transfer function in cryo-electron microscopy in an attempt to improve the fidelity of reconstructed electron-microscope models (Thuman-Commike *et al.*, 1999). A low-resolution structure model derived from solution X-ray scattering and electron microscopy was used to obtain initial phases for determining the crystal structure of *E. coli* ClpP, a molecular chaperone (Wang *et al.*, 1997).

19.3.3.4.2. Time-resolved studies

The advent of synchrotron-radiation sources made time-resolved X-ray studies of biological macromolecular systems possible. Time-resolved small-angle scattering studies on noncrystalline systems, such as skeletal muscle fibres and proteins in solution, are among the first studies of this type. There is still no other structural technique that offers the means of studying large, real-time conformational changes of macromolecular complexes that cannot be contained within a crystal lattice. This aspect of time-resolved solution scattering thus complements time-resolved Laue crystallography. Arguably, the most important contribution of small-angle solution scattering in recent years is the study of temporal changes in biological molecules resulting from changes in the solution environment. These changes are frequently induced by rapid mixing using a stopped-flow apparatus or a temperature jump that allows a virtually synchronous initiation of a reaction, so that the data reflect a common state for the vast majority of the molecules in solution. The changes may be associations, dissociations [*e.g.* the self-assembly of tobacco mosaic virus (Potschka *et al.*, 1988) and microtubule assembly and oscillation behaviours (Mandelkow *et al.* 1989)], as well as changes in quaternary structure. In time-resolved studies of the allosteric transition of *E. coli* aspartate transcarbamoylase (Tsuruta *et al.*, 1994), a structural intermediate during the enzyme reaction that differs from representative T and R quaternary structures was sought. As the technique has matured, a number of challenging studies have recently been conducted. Many of these time-resolved studies are focused on protein folding, as mentioned below.

19.3.3.4.3. Protein-folding studies

There have been a number of solution X-ray scattering studies on protein folding in recent years, in which proteins of low molecular weight had to be investigated under low electron-density contrast due to the presence of denaturing agents, such as urea, at high concentration. Solution X-ray scattering complements other structural techniques used in protein-folding studies, as this is the only technique available for learning how compact a protein is in solution. A Kratky plot of solution scattering data serves as a quick means of determining whether a protein is folded or unfolded. Recently, the singular value decomposition method was applied to this class of problems and revealed a folding intermediate of lysozyme (Chen *et al.*, 1996). This study has recently been expanded to a time-resolved study, which revealed a compact folding intermediate that has not yet formed a hydrophobic core (Chen *et al.*, 1998). Similar studies are being carried out on other protein systems. Arai *et al.* (1998) reported that β-lactoglobulin undergoes a similar folding pattern while it forms a folding intermediate with a hydrophobic core within 100 ms. Uversky *et al.* (1998) reported association-induced folding of globular proteins. Pollack *et al.* (1999) developed a micro-machined mixer to study folding of cytochrome *c* in the sub-millisecond regime.

References

Arai, M., Ikura, T., Semisotnov, G. V., Kihara, H., Amemiya, Y. & Kuwajima, K. (1998). *Kinetic refolding of β-lactoglobulin. Studies by synchrotron X-ray scattering, and circular dichroism, absorption and fluorescence spectroscopy*. *J. Mol. Biol.* **275**, 149–162.

Bonneté, F., Malfois, M., Finet, S., Tardieu, A., Lafont, S. & Veesler, S. (1997). *Different tools to study interaction potentials in γ-crystallin solutions: relevance to crystal growth*. *Acta Cryst.* D**53**, 438–447.

Boulin, C. J., Kempf, A., Gabriel, A. & Koch, M. H. J. (1988). *Data acquisition systems for linear and area X-ray detectors using delay line readout*. *Nucl. Instrum. Methods Phys. Res. A*, **269**, 312–320.

Bu, Z., Perlo, A., Johnson, G. E., Olack, G., Engelman, D. M. & Wyckoff, H. W. (1998). *A small-angle X-ray scattering apparatus for studying biological macromolecules in solution*. *J. Appl. Cryst.* **31**, 533–543.

Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm*. *Biophys. J.* **74**, 2760–2775.

Chen, L., Hodgson, K. O. & Doniach, S. (1996). *A lysozyme folding intermediate revealed by solution X-ray scattering*. *J. Mol. Biol.* **261**, 658–671.

Chen, L., Wildegger, G., Kiefhaber, T., Hodgson, K. O. & Doniach, S. (1998). *Kinetics of lysozyme refolding: structural characterization of a non-specifically collapsed state using time-resolved X-ray scattering*. *J. Mol. Biol.* **276**, 225–237.

Cipriani, F., Gabriel, A. & Koch, M. H. J. (1994). *Alternative approaches to delay line readout for multiwire proportional chambers*. *Nucl. Instrum. Methods A*, **346**, 286–291.

Czeslik, C., Malessa, R., Winter, R. & Rapp, G. (1996). *High pressure synchrotron X-ray diffraction studies of biological molecules using the diamond anvil technique*. *Nucl. Instrum. Methods Phys. Res. A*, **368**, 847–851.

Dubuisson, J.-M., Decamps, T. & Vachette, P. (1997). *Improved signal-to-background ratio in small-angle X-ray scattering experiments with synchrotron radiation using an evacuated cell for solutions*. *J. Appl. Cryst.* **30**, 49–54.

Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering*. New York: Plenum Press.

Genick, U., Borgstahl, G. E. O., Ng, K., Ren, Z., Pradervand, C., Burke, P. M., Srajer, V., Teng, T.-Y., Schildkamp, W., McRee, D. E., Moffat, K. & Getzoff, E. D. (1997). *Structure of a protein photocycle intermediate by millisecond time-resolved crystallography*. *Science*, **275**, 1471–1475.

Glatter, O. (2004). *X-ray techniques*. In *International Tables for Crystallography*, Vol. C. *Mathematical, Physical and Chemical Tables*, edited by E. Prince, Section 2.6.1. Dordrecht: Kluwer Academic Publishers.

Glatter, O. & Kratky, O. (1982). Editors. *Small Angle X-ray Scattering*. London: Academic Press.

Griffith, J. P., Griffith, D. L., Rayment, I., Murakami, W. T. & Caspar, D. L. D. (1992). *Inside polyomavirus at 25-Å resolution*. *Nature (London)*, **355**, 652–654.

Grossman, J. G., Hasnain, S. S., Yousafzai, F. K., Smith, B. E. & Eady, R. R. (1997). *The first glimpse of a complex of nitrogenase component proteins by solution X-ray scattering: conformation of the electron transfer transition state complex of Klebsiella pneumonia nitrogenase*. *J. Mol. Biol.* **266**, 642–648.

Guinier, A. & Fournet, G. (1955). *Small-Angle Scattering of X-rays*. New York: John Wiley & Sons.

Harrison, S. C. (1969). *Structure of tomato bushy stunt virus. I. The spherically averaged electron density*. *J. Mol. Biol.* **42**, 457–483.

Heidorn, D. B. & Trewhella, J. (1988). *Comparison of the crystal and solution structures of calmodulin and troponin C*. *Biochemistry*, **27**, 909–915.

Henderson, S. J. (1995). *Comparison of parasitic scattering from window materials used for small-angle X-ray scattering: a better beryllium window*. *J. Appl. Cryst.* **28**, 820–826.

Hiragi, Y., Nakatani, H., Kajiwara, K., Inoue, H., Sano, Y. & Kataoka, M. (1988). *Temperature-jump apparatus and measuring system for synchrotron solution X-ray scattering experiments*. *Rev. Sci. Instrum.* **59**, 64–66.

Improta, S., Krueger, J. K., Gautel, M., Atkinson, R. A., Lefevre, J. F., Moulton, S., Trewhella, J. & Pastore, A. (1998). *The assembly of immunoglobulin-like modules in titin: implications for muscle elasticity*. *J Mol. Biol.* **284**, 761–777.

Inouye, H., Tsuruta, H., Sedzik, J., Uyemura, K. & Kirschner, D. A. (1999). *Tetrameric assembly of full-sequence protein zero myelin glycoprotein by synchrotron X-ray scattering*. *Biophys. J.* **76**, 423–437.

Jack, A. & Harrison, S. C. (1975). *On the interpretation of small-angle X-ray solution scattering from spherical viruses*. *J. Mol. Biol.* **99**, 15–25.

Johnson, J. E. & Hollingshead, C. (1981). *Crystallographic studies of cowpea mosaic virus by electron microscopy and X-ray diffraction*. *J. Ultrastruct. Res.* **74**, 223–231.

Kataoka, M., Head, J. F., Seaton, B. A. & Engelman, D. M. (1989). *Melittin binding causes a large calcium-dependent conformational change in calmodulin*. *Proc. Natl Acad. Sci. USA*, **86**, 6944–6948.

Kihara, H. (1994). *Stopped-flow apparatus for X-ray scattering and XAFS*. *J. Synchrotron Rad.* **1**, 74–77.

Koch, M. H. J. (1988). *Instruments and methods for small-angle scattering with synchrotron radiation*. *Makromol. Chem. Macromol. Symp.* **15**, 79–90.

Koch, M. H. J. (1991). *Scattering from non-crystalline systems*. In *Handbook on Synchrotron Radiation*, Vol. 4, edited by S. Ebashi, M. Koch & E. Rubenstein. Amsterdam: Elsevier Science Publishers.

Koch, M. H. J. & Svergun, D. I. (1992). Unpublished result.

König, S., Svergun, D., Koch, M. H. J., Höbner, G. & Schellenberger, A. (1992). *Synchrotron radiation solution X-ray scattering study of the pH dependence of the quaternary structure of yeast pyruvate decarboxylase*. *Biochemistry*, **31**, 8726–8731.

Kozin, M. B., Volkov, V. V. & Svergun, D. I. (1997). *ASSA, a program for three-dimensional rendering in solution scattering from biopolymers*. *J. Appl. Cryst.* **30**, 811–815.

Krueger, J. K., Glah, G. A., Rokop, S. E., Zhi, G., Stull, J. T. & Trewhella, J. (1997). *Structures of calmodulin and a functional myosin light chain kinase in the activated complex: a neutron scattering study*. *Biochemistry*, **36**, 6017–6023.

Lattman, E. E. (1989). *Rapid calculation of the solution scattering profile from a macromolecule of known structure*. *Protein Struct. Funct. Genet.* **5**, 149–155.

Mandelkow, E., Lange, G. & Mandelkow, E. M. (1989). *Applications of synchrotron radiation to the study of biopolymers in solution time-resolved X-ray scattering of microtubule self-assembly and oscillations*. *Top. Curr. Chem.* **151**, 2–30.

Miller, S. T., Genova, J. D. & Hogle, J. M. (1999). *Collection of very low resolution protein data*. *J. Appl. Cryst.* **32**, 1183–1185.

Moffat, K. (1997). *Laue diffraction*. *Methods Enzymol.* **277**, 433–447.

Moore, P. B. (1980). *Small-angle scattering. Information content and error analysis*. *J. Appl. Cryst.* **13**, 168–175.

Mourey, L., Pédelacq, J.-D., Fabre, C., Causse, H., Rougé, P. & Samama, J.-P. (1997). *Small-angle X-ray scattering and crystallographic studies of arcelin-1: an insecticidal lectin-like glycoprotein from Phaseolus vulgaris L*. *Proteins*, **29**, 433–442.

Olah, G. A., Gray, D. M., Gray, C. W., Kergil, D. L., Sosnick, T. R., Mark, B. L., Vaughan, M. R. & Trewhella, J. (1995). *Structures of fd gene 5 protein–nucleic acid complexes: a combined solution scattering and electron microscopy study*. *J. Mol. Biol.* **249**, 576–594.

Petitpas, I., Lepault, J., Vachette, P., Charpilienne, A., Mathieu, M., Kohli, E., Pothier, P., Cohen, J. & Rey, F. A. (1998). *Crystallization and preliminary X-ray analysis of rotavirus protein VP6*. *J. Virol.* **72**, 7615–7619.

Petrascu, A.-M., Koch, M. H. J. & Gabriel, A. (1998). *A beginners' guide to gas-filled proportional detectors with delay line readout. J. Macromol. Sci. Phys.* B**37**, 463–483.

Pilz, I., Glatter, O. & Kratky, O. (1979). *Small-angle X-ray scattering. Methods Enzymol.* **61**, 148–249.

Pollack, L., Tate, M. W., Darnton, N. C., Knight, J. B., Gruner, S. M., Eaton, W. A. & Austin, R. H. (1999). *Compactness of the denatured state of a fast-folding protein measured by submillisecond small-angle X-ray scattering. Proc. Natl Acad. Sci. USA*, **96**, 10115–10117.

Potschka, M., Koch, M. H. J., Adams, M. L. & Schuster, T. M. (1988). *Time-resolved solution X-ray scattering of tobacco mosaic virus coat protein: kinetics and structure of intermediates. Biochemistry*, **27**, 8481–8491.

Schindelin, H., Kisker, C., Schlessman, J. L., Howard, J. B. & Rees, D. C. (1997). *Structure of ADP. AIF-4-stabilized nitrogenase complex and its implications for signal transduction. Nature (London)*, **387**, 370–376.

Schmidt, T., Johnson, J. E. & Phillips, W. (1983). *The spherically averaged structures of cowpea mosaic virus components by X-ray solution scattering. Virology*, **127**, 65–73.

Schuster, M. & Göbel, H. (1995). *Parallel-beam coupling into channel-cut monochromators using curved graded multilayers. J. Phys. D*, **28**, A270–A275.

Seaton, B. A., Head, J. F., Engelman, D. M. & Richards, F. M. (1985). *Calcium-induced increase in the radius of gyration and maximum dimension of calmodulin measured by small-angle X-ray scattering. Biochemistry*, **24**, 6740–6743.

Sousa, M. C. & McKay, D. B. (1998). *The hydroxyl of threonine 13 of the bovine 70-kDa heat shock cognate protein is essential for transducing the ATP-induced conformational change. Biochemistry*, **37**, 15392–15399.

Srajer, V., Teng, T.-Y., Ursby, T., Pradervand, C., Ren, Z., Adachi, S., Bourgeois, D., Wulff, M. & Moffat, K. (1996). *Photolysis of the carbon monoxide complex of myoglobin: nanosecond time-resolved crystallography* [see *Comments*]. *Science*, **274**, 1726.

Stuhrmann, H. B. (1970). *Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle-scattering function. Acta Cryst.* A**26**, 297–306.

Sunnerhagen, M., Olah, G. A., Stenflo, J., Forsen, S., Drakenberg, T. & Trewhella, J. (1996). *The relative orientation of Gla and EGF domains in coagulation factor X is altered by Ca-2+ binding to the first EGF domain. A combined NMR–small angle X-ray scattering study. Biochemistry*, **35**, 11547–11559.

Svergun, D. I. (1992). *Determination of the regularization parameter in indirect-transform methods using perceptual criteria. J. Appl. Cryst.* **25**, 495–503.

Svergun, D. I. (1999). *Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys. J.* **76**, 2879–2886.

Svergun, D. I., Aldag, I., Sieck, T., Altendorf, K., Koch, M. H. J., Kane, D. J., Kozin, M. B. & Grueber, G. (1998). *A model of the quaternary structure of the Escherichia coli F1 ATPase from X-ray solution scattering and evidence for structural changes in the Delta subunit during ATP hydrolysis. Biophys. J.* **75**, 2212–2219.

Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Cryst.* **28**, 768–773.

Svergun, D. I., Barberato, C., Koch, M. H. J., Fetler, L. & Vachette, P. (1997). *Large differences are observed between the crystal and solution quaternary structures of allosteric aspartate transcarbamylase in the R state. Proteins Struct. Funct. Genet.* **27**, 110–117.

Svergun, D. I., Koch, M. H. & Serdyuk, I. N. (1994). *Structural model of the 50 S subunit of Escherichia coli ribosomes from solution scattering. I. X-ray synchrotron radiation study. J. Mol. Biol.* **240**, 66–77.

Svergun, D. I., Konrad, S., Huss, M., Koch, M. H. J., Wieczorek, H., Altendorf, K., Volkov, V. V. & Grueber, G. (1998). *Quaternary structure of V_1 and F_1 ATPase: significance of structural homologies and diversities. Biochemistry*, **37**, 17659–17663.

Svergun, D. I., Richard, S., Koch, M. H. J., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Protein hydration in solution: experimental observation by X-ray and neutron scattering. Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *New developments in direct shape determination from small-angle scattering. 2. Uniqueness. Acta Cryst.* A**52**, 419–426.

Svergun, D. I., Volkov, V. V., Kozin, M. B., Stuhrmann, H. B., Barberato, C. & Koch, M. H. J. (1997). *Shape determination from solution scattering of biopolymers. J. Appl. Cryst.* **30**, 798–802.

Thuman-Commike, P. A., Tsuruta, H., Greene, B., Prevelige, P. E., King, J. & Chiu, W. (1999). *Solution X-ray scattering based estimation of electron cryomicroscopy imaging parameters for reconstruction of virus particles. Biophys. J.* **76**, 2249–2261.

Trewhella, J. (1998). *Insights into biomolecular function from small-angle scattering. Curr. Opin. Struct. Biol.* **7**, 702–708.

Tsuruta, H., Reddy, V., Wikoff, W. & Johnson, J. (1998). *Imaging RNA and dynamic protein segments with low resolution virus crystallography; experimental design, data processing and implications of electron density maps. J. Mol. Biol.* **284**, 1439–1452.

Tsuruta, H., Vachette, P., Sano, T., Moody, M. F., Amemiya, Y., Wakabayashi, K. & Kihara, H. (1994). *Kinetics of the quaternary structure change of aspartate transcarbamylase triggered by succinate, a competitive inhibitor. Biochemistry*, **33**, 10007–10012.

Uversky, V. N., Segel, D. J., Doniach, S. & Fink, A. L. (1998). *Association-induced folding of globular proteins. Proc. Natl Acad. Sci. USA*, **95**, 5480–5483.

Wang, J., Harting, J. A. & Flanagan, J. M. (1997). *The structure of ClpP at 2.3 Å resolution suggests a model for ATP-dependent proteolysis. Cell*, **91**, 447–456.

Weik, M., Ravelli, R. B. G., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Specific chemical and structural damage to proteins produced by synchrotron radiation. Proc. Natl Acad. Sci. USA*, **97**, 623–628.

Wikoff, W., Duda, R., Hendrix, R. & Johnson, J. (1998). *Crystallization and preliminary X-ray analysis of the dsDNA bacteriophage HK97 mature empty capsid. Virology*, **243**, 113–118.

Wilbanks, S. M., Chen, L., Tsuruta, H., Hodgson, K. O. & McKay, D. B. (1995). *Solution small-angle X-ray scattering study of the molecular chaperone Hsc70 and its subfragments. Biochemistry*, **34**, 12095–12106.

Yamashita, M. M., Almassy, R. J., Janson, C. A., Cascio, D. & Eisenberg, D. (1989). *Refined atomic model of glutamine synthetase at 3.5 Å resolution. J. Biol. Chem.* **264**, 17681–17690.

Zheng, Y., Doerschuk, P. C. & Johnson, J. E. (1995). *Determination of three-dimensional low-resolution viral structure from solution X-ray scattering data. Biophys. J.* **69**, 619–639.

International Tables for Crystallography (2012). Vol. F, Chapter 19.4, pp. 575–582.

Chapter 19.4. Small-angle neutron scattering

D. M. Engelman and P. B. Moore

19.4.1. Introduction

Diffuse scatter results when X-ray and neutron beams pass through gases and liquids. It is caused by local inhomogeneities, which all fluids invariably contain, at least transiently, and information about them can be obtained by analysis of the scatter they cause. This diffuse scatter is rotationally symmetric about the direction defined by the incident beam because, on average, gases and liquids are isotropic, and it does depend on scattering angle, 2θ. The intensity of the diffuse X-ray scatter of water, for example, is small at small 2θ, and it reaches a maximum at equivalent Bragg spacings equal to the reciprocal of the average oxygen–oxygen distance. In addition to a 'water ring' at high scattering angles, the diffuse scatter of macromolecular solutions includes a peak at $2\theta = 0$, due to the presence of the macromolecules themselves. If ångstrom-wavelength radiation is used, the central macromolecular peak is entirely contained in the region where $\sin(2\theta) \simeq 2\theta$. This is the region examined in small-angle scattering experiments.

Several properties of macromolecules can be determined by analysing their small-angle solution scattering, among them molecular weight, radius of gyration and maximum linear dimension. Approximate shapes can sometimes be obtained, and if a macromolecule is a complex of different chemical species, information about the distribution of its components may emerge. Hydration and conformational changes are also studied this way.

The molecular properties that can be investigated by small-angle scattering are the same for thermal neutrons and X-rays, but the advantages of neutrons are so great that if the equipment required were not so expensive, not many would do small-angle X-ray scattering (SAXS). They are all manifestations of the differences between the ways in which neutrons and X-ray photons interact with matter. For example, thermal neutrons have very low kinetic energies ($\simeq kT$), and, consequently, the energy they deposit in a sample when they are scattered inelastically is negligible. X-ray photons have large energies, and when they are absorbed or scattered inelastically, damaging amounts of energy are deposited. Thus, samples are 'safer' in neutron beams than they are in X-ray beams.

The cross section for X-ray absorption rises so fast with increasing wavelength that it is impractical to do solution-scattering experiments using X-rays with wavelengths much greater than 1.5 Å. The combination of this and the fact that X-ray beams scatter strongly off the edges of optical-track components makes it difficult to build small-angle X-ray spectrometers that measure diffuse scatter at equivalent reciprocal spacings of 0.001 Å^{-1} or less. The cross section for thermal neutron absorption is small and nearly independent of wavelength over the range 1–10 Å. Furthermore, parasitic neutron scatter is easy to control. Thus, it is comparatively straightforward to build small-angle neutron scattering (SANS) spectrometers that measure diffuse scatter at reciprocal spacings considerably less than 0.001 Å^{-1}.

Even more important to those interested in SANS are the vistas opened up by the huge difference in scattering length that exists between ^{1}H and ^{2}H (henceforth termed H and D), to which we will return below. In brief, the scatter of macromolecular solutions can be significantly altered by replacing some or all of the H atoms with D atoms. This control greatly extends the range of problems that can be addressed by SANS, and the chemical 'cost' is minimal. A perdeuterated molecule is almost identical to its protonated counterpart. The X-ray scattering of substances depends on the number of electrons they contain, and when this number is changed, chemical properties change also.

Relative to SAXS, the sole disadvantage of SANS is a phenomenon called incoherent scatter, which is a comparatively minor aspect of X-ray work. Neutrons are scattered primarily by atomic nuclei, and if a nucleus has spin, its scattering length depends on the orientation of its spin relative to that of each neutron with which it interacts. Since nuclear spins are usually unoriented in SANS samples, this spin-orientation dependence leads to a random atom-to-atom variation in scattering length. Coherent scatter, on which all diffraction effects depend, is determined by average scattering-length values. The scatter due to fluctuations about the average is incoherent, and in the low-angle region incoherent scatter manifests itself as a featureless background that is independent of scattering angle. The cross section for incoherent scattering is very large for H atoms, and since both water and biological macromolecules contain large proportions of H atoms, incoherent scatter is often a dominant source of background.

Some useful general references for small-angle scattering in general and neutron scattering in particular are Bacon (1975), Glatter & Kratky (1982) and Guinier (1955, 1962).

19.4.2. Fundamental relationships

For most purposes, a dilute macromolecular solution can be thought of as a macromolecular gas, and for that reason it is appropriate to apply Debye's theory for gas scatter to macromolecular solutions (Debye, 1915). Debye's master equation can be cast into neutron terms as follows:

$$I(Q) \propto I_0 \sum \sum b_i b_j \sin(Qr_{ij})/(Qr_{ij}), \qquad (19.4.2.1)$$

where $I(Q)$ is the amount of scattered radiation observed at Q, I_0 is the intensity of the incident beam, b_i and b_j are the scattering lengths of the ith and jth atoms in the molecule, r_{ij} is the distance between atoms i and j, and $Q = (4\pi/\lambda)\sin\theta$, λ being the wavelength of the radiation used and θ being half the scattering angle. When applied to molecules in solution, both summations must include not only all the atoms that are covalent components of the molecule in question, but also all the associated solvent atoms, because when a macromolecule dissolves, the inhomogeneity created includes the counterions associated with it, its solvation layer *etc.* Equation (19.4.2.1) holds for a single molecule; if the number of molecules contributing to scattering in some sample is N, the scattering profile measured will be N times the profile due to a single molecule. The inhomogeneities responsible for small-angle scatter have linear dimensions of the order of 10 Å or more and, hence, have volumes that contain large numbers of atoms. In addition, interatomic spacings cannot

Copyright © 2012 International Union of Crystallography

be resolved using small-angle data. Thus, it is appropriate to discuss small-angle scattering in terms of electron densities for X-rays or scattering-length densities for thermal neutrons. The scattering-length density of a volume, ρ, is given by

$$\rho \equiv \textstyle\sum b_i / V, \tag{19.4.2.2}$$

where b_i is the scattering length of the ith atom in volume V and the summation runs over all atoms in the volume.

Recasting the Debye equation in terms of scattering lengths, one obtains

$$I(Q) \propto I_0 \textstyle\int \int \rho(\mathbf{r}_i)\rho(\mathbf{r}_j)[\sin(Qr_{ij})/(Qr_{ij})] \, \mathrm{d}V_i \, \mathrm{d}V_j,$$

where $\rho(\mathbf{r}_i)$ and $\rho(\mathbf{r}_j)$ are the scattering-length densities in volume elements whose positions are described by vectors $\mathbf{r}_i$ and $\mathbf{r}_j$, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, and both integrals run over the volume of the entire macromolecule plus the surrounding perturbed volume. Both this equation and equation (19.4.2.1) implicitly assume that the medium surrounding each molecule is a vacuum, which is not true for the molecules in solution. The effect of solvent on low-angle scattering can be taken care of by subtracting the average scattering length of the solvent, ρ_0, from the scattering-length densities within the molecule. Thus,

$$I(Q) \propto I_0 \textstyle\int \int [\rho(\mathbf{r}_i) - \rho_0][\rho(\mathbf{r}_j) - \rho_0]_j [\sin(Qr_{ij})/(Qr_{ij})] \, \mathrm{d}V_i \, \mathrm{d}V_j. \tag{19.4.2.3}$$

The quantity $[\rho(\mathbf{r}_i) - \rho_0]$ is a *contrast*, and, as will be shown, contrast manipulation is a major component of SANS experiments.

Equation (19.4.2.3) can be evaluated a second way, because the $(\sin x)/x$ term in the integral depends only on the distances between volume elements, not on their locations in space. Thus, a function $p(r)$ can be defined as follows:

$$p(r) \equiv \textstyle\int \int [\rho(\mathbf{r}_i) - \rho_0][\rho(\mathbf{r}_i + \mathbf{r}) - \rho_0] \, \mathrm{d}V_i \, \mathrm{d}V\mathbf{r},$$

where the integral in $\mathbf{r}$ runs over all $\mathbf{r}$ such that $|\mathbf{r}| = r$, and the integral in $\mathbf{r}_i$ runs over the entire molecular volume. Written in terms of $p(r)$, equation (19.4.2.3) becomes

$$I(Q) \propto I_0 \textstyle\int p(r)[\sin(Qr)/(Qr)] \, \mathrm{d}r, \tag{19.4.2.4}$$

where the integral runs from $r = 0$ to $r_{\max}$, the maximum atom-to-atom length within the molecule.

Note that if contrast was constant within a macromolecule, $p(r)$ would be proportional to the distribution of interatomic distances in the molecule, and for that reason $p(r)$ is often called the *length distribution*. Note also that $p(r)$ is simply the molecule's Patterson function, rotationally averaged about its origin. Note, finally, that $p(r)$ is the summation of a large, but finite, number of sharp, discrete interatomic distance peaks, each with its own weight. If the individual interatomic peaks in this 'length spectrum' could be assigned, *i.e.*, if the atoms responsible for each one could be identified, it would be possible to determine the three-dimensional structure of the molecule in question, save for uncertainty about its hand.

Since solution-scattering profiles can be computed by sine transformations of length distributions, it is reasonable to hope that a transformation might exist that enables one to compute length distributions once solution-scattering profiles have been measured. There is (Debye & Bueche, 1949; Debye & Pirenne, 1938):

$$p(r) \propto r \textstyle\int QI(Q) \sin(Qr) \, \mathrm{d}r. \tag{19.4.2.5}$$

Two practical issues must be addressed when carrying out the operation implied by equation (19.4.2.5) because the integral it contains runs from $Q = 0$ to ∞. Firstly, scattering is never measured at $Q = 0$ due to interference with the direct beam. Secondly, the largest value of Q for which $I(Q)$ is measured is always less than ∞. The absence of data at very small values of Q is easily addressed, because a soundly based method exists for extrapolating the low-angle data to $Q = 0$ (see below). The lack of data at high Q is harder to cope with, but it can be dealt with approximately using Porod's Law (Porod, 1951, 1952) and the impact of its absence on molecular parameters deduced from small-angle data is easy to estimate. In any case, it is important to realize that length distributions represent the sum total of the information that can be extracted from solution-scattering experiments.

The problem of extrapolating small-angle data to $Q = 0$ was solved by Guinier (1939). He demonstrated that, at very small angles,

$$I(Q) \propto I(0) \exp[-(QR_g)^2/3], \tag{19.4.2.6}$$

where R_g is the radius of gyration, and

$$R_g \equiv (\{\textstyle\int [\rho(\mathbf{r}) - \rho_0]|\mathbf{r}|^2 \, \mathrm{d}V\}/\{\textstyle\int [\rho(\mathbf{r}) - \rho_0] \, \mathrm{d}V\})^{1/2}. \tag{19.4.2.7}$$

The origin of the vector $\mathbf{r}$ in this equation is the centre of gravity of the macromolecule's scattering-length density distribution, *i.e.*, it is the point where

$$0 = \{\textstyle\int [\rho(\mathbf{r}) - \rho_0]\mathbf{r} \, \mathrm{d}V\}/\{\textstyle\int [\rho(\mathbf{r}) - \rho_0] \, \mathrm{d}V\}.$$

It follows from equation (19.4.2.6) that if the lowest-angle data collected are plotted in the form $\ln[I(Q)]$ versus Q^2, a straight line should result, the slope of which is $(R_g^2/3)$ and the intercept of which at $Q = 0$ is $I(0)$. Note that data have to be obtained at scattering angles well inside the region where $I(Q) \sim I(0)/2$ in order for this formula to hold; if the data are thus obtained, a radius of gyration estimate will emerge. The radius of gyration of an object is the root-mean-squared distance between its centre of gravity and the elements of which it is composed.

As might be expected, $I(0)$ and R_g can also be computed from $p(r)$. Consider the magnitude of $I(Q)$ at $Q = 0$. Since the $\sin x/x$ term in equation (19.4.2.4) is 1 at $Q = 0$,

$$I(0) \propto I_0 \textstyle\int \int [\rho(\mathbf{r}_i) - \rho_0][\rho(\mathbf{r}_j) - \rho_0] \, \mathrm{d}V_i \, \mathrm{d}V_j = \textstyle\int p(r) \, \mathrm{d}r. \tag{19.4.2.8}$$

Thus, $I(0)$, the forward scatter, is proportional to the integral of the length distribution. It is easy to show that R_g equals $(M/2)^{1/2}$, where M is the second moment of $p(r)$ given by

$$M = [\textstyle\int r^2 p(r) \, \mathrm{d}r]/[\textstyle\int p(r) \, \mathrm{d}r]. \tag{19.4.2.9}$$

The average atom-to-atom distance in a molecule, $r_{\rm ave}$, is easy to compute if $p(r)$ is known from

$$r_{\rm ave} = [\textstyle\int rp(r) \, \mathrm{d}r]/[\textstyle\int p(r) \, \mathrm{d}r]. \tag{19.4.2.10}$$

The reason forward scatter, $I(0)$, is interesting is its dependence on molecular weight. As equation (19.4.2.8) suggests, the forward scatter measured for a sample is proportional to N times the square of the product of the average contrast between a molecule and its solvent and the molecular volume, where N is again the number of molecules contributing to the signal observed. Since average contrasts can be estimated from chemical compositions and partial specific volumes, $I(0)$ measurements can be used to estimate molecular weights. If the $I(0)$ values of solutions of a set of molecules of similar chemical composition are compared, it will be found that $I(0)$ divided by the weight concentration of each sample is proportional to molecular weight.

This procedure for estimating molecular weights can fail. Suppose $\int[\rho(\mathbf{r}_i) - \rho_0]\, \mathrm{d}V = 0$, *i.e.*, the scattering-length density of the solvent is the same as the average scattering-length density of the macromolecule. Then $I(0)$ will be zero, the solution-scattering profile will lack a peak at small angles and no molecular-weight estimate will result. Under these conditions, the macromolecule is said to be 'contrast matched'. It is easy to contrast-match biological macromolecules in the context of SANS experiments, because all biological macromolecules that have not been labelled with ^{2}H have average scattering-length densities between those of H_2O and D_2O (see below).

19.4.3. Contrast variation

19.4.3.1. Variation of solvent density

The principle of contrast variation was studied in early work by Bragg & Perutz (1952), who observed that the magnitudes of low-order reflections in X-ray studies of protein crystals were reduced as the salt concentration in the solvent was raised. Following their concept, the effective scattering density of a dissolved particle is

$$\rho(\mathbf{r}) = \rho(\mathbf{r})_{\text{solute}} - \rho(\mathbf{r})_{\text{solvent}} = \rho(\mathbf{r})_{\text{solute}} - \rho_{\text{solvent}},$$

where ρ_{solvent} is the average solvent scattering density. If we take a mean scattering density for the particle, a mean contrast, ρ, is defined:

$$\rho = \rho_{\text{solute}} - \rho_{\text{solvent}}. \qquad (19.4.3.1)$$

If the solute and solvent have equal densities, they are 'contrast matched', and the scattering from the particle vanishes at zero angle. The particle will nonetheless scatter radiation at larger angles as a consequence of internal density fluctuations, which can be useful in obtaining structural information. In the case of neutron scattering, the solvent density is most often varied by mixing D_2O and H_2O to obtain different percentages of D_2O. This strategy gives a range of solvent densities that includes the densities of most biological molecules. However, biological molecules contain hydrogen atoms that exchange with solvent, so deuteration of the environment alters their scattering density to some extent (see below). In general, all hydrogen atoms not bonded to carbon are potentially exchangeable, but not all of these will actually exchange in a typical experiment.

To describe the variation of the radius of gyration with contrast, Stuhrmann derived the useful relationship (Stuhrmann, 1976; Stuhrmann *et al.*, 1976)

$$R_g^2 = R_c^2 + \alpha/\rho + \beta/\rho^2, \qquad (19.4.3.2)$$

which separates the contributions of the internal structure of the particle (ρ_I) to its radius of gyration from the contributions of the shape (ρ_c). Scattering from the internal structure is independent of contrast; scattering from the shape is contrast dependent. The shape function is defined as having a value of one inside the particle and zero outside. The total scattering density is then

$$\rho(\mathbf{r}) = \rho\rho_c(\mathbf{r}) + \rho_I(\mathbf{r}). \qquad (19.4.3.3)$$

The contrast-independent terms in the Stuhrmann equation are

$$\alpha = (1/V_c) \textstyle\int \rho_I(\mathbf{r}) r^2 \, \mathrm{d}^3 r, \qquad (19.4.3.4)$$

$$\beta = (1/V_c^2) \textstyle\int\int \rho_I(\mathbf{r})\rho_I(\mathbf{r}')\mathbf{r}\mathbf{r}' \, \mathrm{d}^3 r \, \mathrm{d}^3 r' \text{ and} \qquad (19.4.3.5)$$

$$R_c^2 = (1/V_c) \textstyle\int \rho_c(\mathbf{r}) r^2 \, \mathrm{d}^3 r, \qquad (19.4.3.6)$$

where R_c is the radius of gyration of the shape function and V_c is its volume. The sign and magnitude of α give information on the radial density distribution of scattering density in the particle: if the outer region is higher in density than the inner region, α is positive (as, for example, in lipoproteins); if the inner region is denser, α is negative. The β coefficient represents the displacement of the centre of mass as a function of the contrast and is always positive; in real cases, β is often negligible. The Stuhrmann equation leads to a useful way to represent graphically the radius of gyration data obtained from a series of contrasts: the observed R_g^2 is plotted *versus* $1/\rho$. If β is negligible, the plot is a straight line of slope α, intercepting the $1/\rho$ axis at R_c^2. Thus, R_c is obtained by extrapolation to a point where $\rho = \infty$, and so is often termed the radius of gyration at infinite contrast. This quantity is a representation of the shape of the particle as if it had uniform internal scattering density. In a particle with two discrete regions of density, the radius of gyration for each region can be obtained from such a graph by evaluating R_g^2 where ρ is equal to the density of one region, so that R_g^2 of the non-contrast-matched region is determined. Such measurements can also be made by adjusting the solvent to match the scattering of one region to reveal the scattering of the other.

A parameter that is often useful is the contrast-match point for the particle, which reflects its overall composition including exchange.

$$\rho_M = \textstyle\sum b_I/V + nd(b_{\mathrm{D}} - b_{\mathrm{H}})/V, \qquad (19.4.3.7)$$

where ρ_M, the match point, is the solvent scattering length density at which the contrast is zero and n is the number of exchanged hydrogens multiplied by d, the fractional deuteration of the water at the match point. Typically, the match point is obtained by measuring small-angle scattering at a series of D_2O:H_2O ratios, plotting each using a Guinier plot $\{\ln[I(Q) - I(0)]$ *versus* $Q^2\}$ to obtain a value for $I(0)$ by extrapolation, and then plotting $[I(0)/C]^{1/2}$ *versus* ρ_{solvent}, where C is the particle concentration. It is often convenient to represent ρ_{solvent} as per cent D_2O. The plot should be a straight line, passing through zero at the contrast-match point. As noted above, the vast majority of biological molecules have contrast-match points at densities between those of H_2O and D_2O. If the particles are compositionally heterogeneous, the observed plot will be a weighted sum of the curves for each of the compositions present and will deviate from a straight line at low contrasts. Thus, the contrast-matching experiment can provide information on both composition and homogeneity.

While contrast variation is most often based on variation of the deuteration level in water, it is also possible to create variation by adding molecules to the solvent. As an example, a study of hydration layers was conducted by adding solute molecules, such as glycerol, at high concentration; the solute molecules alter the solvent scattering length density but do not penetrate the hydration layer (Lehmann & Zaccai, 1984).

19.4.3.2. Variation of internal contrast

A second form of contrast variation can be achieved by replacing the hydrogen atoms in biological molecules with deuterium. If, in addition, it is possible to deuterate selected regions of a particle, internal contrasts can be modified to gain information about substructures. In some cases, the opportunity arises because significant biochemical differences are present, as between the RNA and protein portions of a ribosome, the DNA and protein of a nucleosome, or the lipid and protein of a lipoprotein particle. While the intrinsic contrast may be sufficient to provide key information [as, for example, the early finding that

DNA is on the outside of nucleosomes (Bradbury *et al.*, 1976; Uberbacher *et al.*, 1982)], it can be accentuated by incorporation of deuterated biochemical precursors or reconstitution from separately labelled components.

Internal contrast can be created in single molecules by differential incorporation of biosynthetic precursors or by chemical synthesis. Differential incorporation was first used to test models of bacteriorhodopsin, using the incorporation of deuterated amino acids supplied to a culture of halophilic archae (Engelman & Zaccai, 1980), and internal labelling was used to document conformations of cholesteryl esters that had been chemically labelled in key positions (see below). Reconstitution from purified components has been used to place deuterated proteins in ribosomes and in other complexes, again with the aim of creating internal contrast to enhance the information obtained in a neutron experiment.

In general, the creation of internal contrast can be viewed as a strategy for enriching the low information content of a solution-scattering experiment by building additional information into the sample. By design, it is known what has been labelled, so the scattering given by contrasting elements provides information about the relationships of the labelled parts to each other or to the particle as a whole. A particularly informative (but difficult) strategy is to use internal contrast to measure distances between locations in a molecule or complex. Such measurements are discussed below.

19.4.3.3. Relationship of contrasting regions

Where a particle has two regions of different scattering density, the square of the total observed radius of gyration, R_T^2, can be obtained as the weighted sum of squares of the two individual radii, R_1^2 and R_2^2, and the square of the distance between the centres of scattering mass of the two regions, d_{12}^2, as

$$R_T^2 = f_1 R_1^2 + f_2 R_2^2 + f_1 f_2 d_{12}^2, \qquad (19.4.3.8)$$

where the f's are the fractions of the total scattering from each of the two regions at the solvent contrast being used in the experiment (Moore *et al.*, 1974). By varying the contrast, a set of differently weighted equations can be obtained, from which the individual radii and the separation can be derived. This method is an alternative to the Stuhrmann analysis described above. An example of the use of this approach, based formally on the parallel axis theorem of mechanics, is found in studies of the ribosome (Moore *et al.*, 1974). An alternative that has proved useful is to combine neutron and X-ray scattering data, since the weighting factors will differ for distinct regions, such as the RNA and protein components of the ribosome (Serdyuk *et al.*, 1979).

19.4.3.4. The triple isotopic substitution method

An innovation in the study of subunits in a reconstituted complex was introduced by Serdyuk *et al.* (1994), who devised a difference method to isolate the scattering from a single subunit. The method requires three particles with different deuteration levels in the subunit: one in which the subunit is heavily deuterated (contrast with the complex = ρ), one in which the subunit is not deuterated (0), and one in which the subunit is deuterated at an intermediate level ($\rho/2$):

$$I_1(Q) = |C|^2,$$
$$I_2(Q) = |C|^2 + 2(\rho/2)F[CS] + (\rho^2/4)|S|^2 \text{ and}$$
$$I_3(Q) = |C|^2 + 2\rho F[CS] + \rho^2|S|^2,$$

where C is the scattering amplitude of the complex, S is the amplitude of the subunit, and $F[CS]$ is the Fourier transform of the correlation function between the complex and the subunit. Scattering is measured (*a*) from an equimolar mixture of complexes with heavily deuterated and non-deuterated subunits and (*b*) from a sample of complexes with subunits with the intermediate level of deuteration. Subtraction of (*b*) from (*a*), weighted so that the two curves are equimolar, gives a net curve for the subunit alone (at half the scattering power that would be seen for a solution of the isolated subunits at the same concentration):

$$I_1(Q) + I_3(Q) - 2I_2(Q) = (\rho^2/2)|S|^2. \qquad (19.4.3.9)$$

The difference curve is not influenced by solvent composition, underlying order, concentration or interparticle interference effects. Thus, at the cost of some difficult biochemistry, the small-angle scattering of a subunit belonging to a large assembly can be observed *in situ*. In practice, the mixture is not equimolar, but is adjusted depending on the intermediate level of deuteration, relaxing some of the difficulty of the biochemistry.

19.4.3.5. Nuclear spin contrast variation

When atomic nuclei of nonzero spin are placed in a magnetic field, the spins orient. If the temperature is near absolute zero, the orientation results in a polarization that is seen by polarized neutrons, resulting in polarization-dependent scattering. Since polarized neutron sources are available, and since biological materials are rich in hydrogen, Stuhrmann has proposed and tested a measurement based on the following idea (Stuhrmann & Nierhaus, 1996). Consider a complex in which all of the hydrogen has been replaced by deuterium except in one subunit or ligand, and prepare a sample that can be frozen to $T < 0.5$ K, placed in a 2.5 T magnetic field and subjected to dynamic spin polarization. Scattering of polarized neutrons is measured twice, once with the hydrogen spins oriented, and once with the spins selectively depolarized using NMR saturation. The difference contains contributions from the hydrogenated region and a cross term between the region and the rest of the complex. Using a modelling approach, Stuhrmann and his colleagues have deduced a structure that locates transfer RNA molecules on a ribosome from polarized neutron data, revealing the promise of this approach (Stuhrmann & Nierhaus, 1996).

19.4.3.6. Interpretation of small-angle scattering using models

There have been many attempts to extract more information from solution-scattering experiments than the radius of gyration and forward scattering, including the distance-measuring strategies discussed below. These attempts are of two kinds: testing models and creating models. Each of these must be cast in the context of the intrinsic information content of a scattering measurement, which can be expressed in terms of the number of independent parameters, n, that can be uniquely extracted from a data set (Moore, 1980).

$$n = Q_{\max} d_{\max}/\pi, \qquad (19.4.3.10)$$

where $Q_{\max}$ is the largest Q at which statistically significant data are measured and $d_{\max}$ is the largest dimension of the particle. A further requirement, normally met in small-angle scattering, is that $Q_{\min} d_{\max}/\pi < 1$.

The information content is a subtle factor in the first class of modelling, where models are tested for agreement with scattering data. Excellent programs have been written for generating predicted scattering curves from atomic coordinates and have been used to explore perturbations between crystal structures and solution organization. A fine example is the work on ATCase by Svergun, Barberato *et al.* (1997); the article also contains references to the programs used.

A more challenging task is to work in the other direction, extracting structural information directly from a scattering curve. Considerable effort has been devoted to work in this area, using approaches based on spherical harmonics, sometimes using sets of spheres to represent structure, and occasionally integrating information from electron microscopy (Svergun, 1994; Svergun, Burkhardt *et al.*, 1997).

19.4.3.7. Use of forward scattering to measure molecular weights

The value of the scattering function at zero scattering angle, which is obtained by extrapolation using a Guinier plot, is related to the molecular weight of the particle. In the neutron small-angle scattering case, the incoherent scattering background from hydrogen provides an internal standard. Using the incoherent background as an absolute calibration of the beam intensity, and knowing the concentration and composition of particles, one can obtain good values for the molecular weight, as pointed out by Jacrot & Zaccai (1981) and Zaccai & Jacrot (1983). This approach applies particularly well to proteins, where the average scattering density does not vary much from case to case, and can provide important data on the stoichiometry of oligomeric complexes. The limit in the accuracy of the measurement arises from limitations in knowing the protein concentration.

19.4.4. Distance measurements

19.4.4.1. Theory and background

By measuring distances between chemically specified points in a macromolecule or macromolecular complex, useful information may be obtained concerning its structure. In the case of a complex with many subunits, successive inter-subunit distance measurements can be combined by triangulation to generate a three-dimensional map specifying the relative positions of subunits and, perhaps, their individual radii of gyration. The only uncertainty is the handedness if more than three distances are combined. A number of approaches exist for making distance measurements, including NMR, fluorescence and electron-microscope techniques in which labelling is employed. Neutron scattering provides a particularly simple approach to such measurements in a conceptual sense. If we consider two labelled centres in a large macromolecule or complex, the interference of the scattering from the two labelled positions, $I_x(Q)$, can be separated from the scattering of the rest of the complex by measuring the scattering of the unlabelled complex, $I(Q)$, the two singly labelled complexes, $I_1(Q)$ and $I_2(Q)$, and the doubly labelled complex, $I_{12}(Q)$, as they are related by

$$I_x(Q) = [I_{12}(Q) + I(Q)] - [I_1(Q) + I_2(Q)]. \tag{19.4.4.1}$$

Using the Debye relationship, it can be shown that the cross term will have the approximate form

$$I_x(Q) = 2f_1 f_2 \sin(d_{12}Q)/d_{12}Q, \tag{19.4.4.2}$$

where the interference intensity results from the correlation between radiation scattered from sites 1 and 2, with scattering strengths f_1 and f_2. It is a damped sinusoidal fringe with a periodicity reciprocally related to the separation between the scattering regions. If the scattering regions are very small compared with the object, there will be nodes at equal intervals in Q; if they are not, $I_x(Q)$ will be a sum of all cross correlations of label positions in the two regions.

19.4.4.2. Neutron distance measurements

In the neutron case, the usual application involves deuterium labelling of positions in the two regions, and the cross term becomes

$$I_x(Q) = 2(b_D - b_H)^2 \sum_i \sum_j (\sin d_{ij}Q)/d_{ij}Q, \tag{19.4.4.3}$$

where the first sum runs over all labelled positions in region 1 and the second over all labelled positions in region 2.

The interference experiment as a technique for studying macromolecules in solution was first proposed in 1947 as an X-ray scattering measurement in which the labels are heavy atoms (Kratky & Worthmann, 1947). It has been discussed several times since and tested (Hoppe, 1972; Vainshtein *et al.*, 1970); however, no biological application has ever been published, since the signals are small (Hoppe, 1972). In 1972, the interference idea was discovered again, but in the context of neutron scattering (Engelman & Moore, 1972). Initially, it was thought that the distance measurement would follow from inspection of the damped sinusoidal term, but it soon became evident that the size and shape of the labelled regions would have an important influence. Using a power series expansion like that originally employed by Guinier, the previous expression becomes (for small Q)

$$I_x(Q) \propto [1 - (1/6)(R_i^2 + R_j^2 + d_{ij}^2)Q^2],$$

where R_i and R_j are the radii of gyration of the subunits *in situ*, and d_{ij} is the separation of their centres of mass. It follows that (Moore *et al.*, 1978; Stöckel *et al.*, 1979)

$$R_g^2 = (1/2)(R_i^2 + R_j^2 + d_{ij}^2). \tag{19.4.4.4}$$

The second moment of a length distribution, M_{ij}, can be related to a radius of gyration. If $p_{ij}(r)$ is the length distribution of the cross term $I_x(Q)$, then

$$M_{ij} = \int r^2 p_{ij}(r)\,\mathrm{d}r = 2R_g^2.$$

Hence (Moore *et al.*, 1978)

$$M_{ij} = R_i^2 + R_j^2 + d_{ij}^2. \tag{19.4.4.5}$$

Thus, M_{ij} is the parameter that contains the information in the difference experiment. If the radii of gyration of the labelled regions are small, then the distance is well measured in a single experiment. This would be the case, for example, if the labels were single heavy atoms in an X-ray experiment. However, in most complexes of macromolecules there will be many pairwise protein–protein relationships where the radii and the separation of centres have comparable magnitudes. One approach to extracting distance information is to know the radii of the subunits *in situ*, by estimation from their molecular weights, by measurement of the isolated subunits or from triple isotopic substitution (see above). The first approach is the least desirable but the easiest, and the last approach is correct but laborious.

If a complex has eight or more distinct subunits, the number of possible measurements of M_{ij}, $n(n-1)/2$, is sufficient to solve for the distances and radii, enriching the information obtained from the experiment. The minimum number of measurements required is $4n - 6$, and for large complexes, such as the ribosome, a great excess of possible measurements exists, permitting refinement of the information through redundancy. Moore &

Weinstein (1979) have described analytical methods for solving the problem with correct error propagation.

19.4.4.3. *The statistical labelling method*

In many biological assemblies, multiple copies of a given subunit are found. In such cases, statistical labelling methods provide useful information. The basic concept is to produce two samples in which labelled subunits are randomly mixed with unlabelled subunits within the particles, and in which the ratios of labelled to unlabelled subunits is different. The difference in the scattering from the two samples will then contain shape and orientation information for the subunits (Fujiwara & Mendelson, 1996). A simple case is that of a dimer in a larger complex. Random introduction of equimolar labelled (D) and unlabelled (H) subunits will result in a mixture of complexes, where HH:HD:DH:DD = 1:1:1:1. A sample can be prepared with an equimolar mixture of all H and all D so that HH:DD = 1:1. The difference in scattering between these samples will be simply the interference cross term for the dimer, but at half the weight it would have in the case where the subunits were distinctly labelled.

19.4.5. Practical considerations

19.4.5.1. *Feasibility*

For a biological scientist, the first question is whether knowledge of the parameters that can be obtained from neutron scattering is of interest for the biological problem under consideration. If the answer is affirmative, the best course of action is to make contact with individuals who have conducted experiments in the past, as well as with biologists at a neutron-scattering facility. As a preliminary guide, a few general points are pertinent.

For a solution-scattering experiment on an unlabelled sample, typical sample volumes are 0.1–0.2 ml at concentrations of around 10 mg ml^{-1}. For contrast variation, samples dialysed to different D_2O levels will be required, so significant amounts of material need to be at hand. Making a set of measurements will require a few hours at a modern facility.

Longer collection times and significantly greater biochemical efforts will be involved for measurements with labelled material, especially if reconstitution is part of the strategy. Labelled biomolecules are most often produced by growth of organisms in D_2O (Moore & Engelman, 1976; Vanatalu *et al.*, 1993), but strategies using chemical synthesis or providing labelled precursors have also been employed. Preparation of samples with labelled ligands is usually more straightforward than the generation of reconstituted complexes, but still requires tests of homogeneity.

19.4.5.2. *Homogeneity and stability*

For measurements to be made, the samples must be transported to a neutron facility and measured for many hours. This requires stability in the usual biochemical sense, and proper experiments will include assays performed after the measurements. Damage from the neutrons themselves is minimal, but handling and transportation impose greater challenges.

Where mixed samples are employed, stability also involves resistance to exchange of subunits or ligands between complexes in the mixture. This is hard to assess in advance, the first sign of trouble often being an absence of an expected difference signal. Given shipment and measuring times, even a slow exchange is significant. With a consequent qualification of the results, cross-linking may be used to stabilize the complex.

In all scattering measurements, homogeneity of the sample is a great advantage in interpretation. However, some variation is tolerable for many purposes, and the level of purity typically sought for crystallization experiments is not usually required.

19.4.5.3. *Solvent conditions*

Background arises from the incoherent scattering of hydrogen. Consequently, the use of high levels of D_2O in the buffer improves the background, but may not be innocuous, since high levels of D_2O have been seen to induce aggregation. In cases where complexes with labelled subunits are measured, it is advantageous to suppress small-angle background scatter by contrast matching the unlabelled regions *via* solvent adjustment.

19.4.6. Examples

19.4.6.1. *Contrast variation*

An early result that changed ideas about nucleosome organization came from measurements relying on the difference in protein and DNA densities (Baldwin *et al.*, 1975; Bradbury *et al.*, 1976). Information on the organization of serum lipoproteins was also based on intrinsic scattering differences (Stuhrmann *et al.*, 1975; Atkinson & Shipley, 1984). The contrast between RNA and protein was used in early ribosome measurements (Crichton *et al.*, 1977; Moore *et al.*, 1974). Recent examples include detergent binding to membrane proteins (Timmins *et al.*, 1991) and the study of membrane protein–lipid complexes (Jeanteur *et al.*, 1994). Elegant use of contrast variation resulted in a structural explanation of the anti-cooperative binding of tRNA to synthetase, which had resisted study by other methods (Dessen *et al.*, 1978).

19.4.6.2. *Contrast matching*

In some applications, use has been made of contrast matching a large part of a complex to study a component with a contrasting scattering density. Examples include studies of ribosomal proteins *in situ* (Nierhaus *et al.*, 1983; Nowotny *et al.*, 1994), proteins of the DNA-dependant RNA polymerase (Stöckel *et al.*, 1980*a*), studies of muscle proteins (Stone *et al.*, 1998) and a view of a membrane protein in a lipid bilayer (Hunt *et al.*, 1997).

19.4.6.3. *Spin contrast variation*

Some applications of the spin contrast variation method have appeared (Junemann *et al.*, 1998; Nierhaus *et al.*, 1998).

19.4.6.4. *Specific deuteration, combination with X-ray measurements*

Many recent measurements use specific labelling and combine information with data from small-angle X-ray scattering. Studies of ligand binding (Bilgin *et al.*, 1998), hydration layers (Svergun *et al.*, 1998), troponin complexes (Olah *et al.*, 1994) and ribosomes (Svergun *et al.*, 1996) are examples that illustrate the approaches.

19.4.6.5. *Distance measurements and triangulation*

By measuring distances and radii of gyration, models of some large complexes have been created. These include a model of the relative positions of all 21 proteins in the small ribosomal subunit from *E. coli* (Capel *et al.*, 1987), a model of seven proteins from the large subunit (May *et al.*, 1992) and a model of the DNA-dependant RNA polymerase (Stöckel *et al.*, 1980*b*). Intramolecular conformational states of cholesterol esters have been observed using distance measurements on molecules chemically labelled in different regions (Burks & Engelman, 1981).

References

Atkinson, D. & Shipley, G. G. (1984). *Structural studies of plasma lipoproteins. Basic Life Sci.* **27**, 211–226.

Bacon, G. E. (1975). *Neutron Diffraction*. Oxford University Press.

Baldwin, J. P., Boseley, P. G., Bradbury, E. M. & Ibel, K. (1975). *The subunit structure of the eukaryotic chromosome. Nature (London)*, **253**, 245–249.

Bilgin, N., Ehrenberg, M., Ebel, C., Zaccai, G., Sayers, Z., Koch, M. H. J., Svergun, D. I., Barberato, C., Volkov, V., Nissen, P. & Nyborg, J. (1998). *Solution structure of the ternary complex between aminoacyl-tRNA, elongation factor Tu, and guanosine triphosphate. Biochemistry*, **37**, 8163–8172.

Bradbury, E. M., Baldwin, J. P., Carpenter, B. G., Hjelm, R. P., Hancock, R. & Ibel, K. (1976). *Neutron-scattering studies of chromatin. Brookhaven Symp Biol.* **27**, IV97–IV117.

Bragg, W. L. & Perutz, M. F. (1952). *The external form of the haemoglobin molecule. I. Acta Cryst.* **5**, 277–283.

Burks, C. & Engelman, D. M. (1981). *Cholesteryl myristate conformation in liquid crystalline mesophases determined by neutron scattering. Proc. Natl Acad. Sci. USA*, **78**, 6863–6867.

Capel, M. S., Engelman, D. M., Freeborn, B. R., Kjeldgaard, M., Langer, J. A., Ramakrishnan, V., Schindler, D. G., Schneider, D. K., Schoenborn, B. P. & Sillers, I. Y. (1987). *A complete mapping of the proteins in the small ribosomal subunit of Escherichia coli. Science*, **238**, 1403–1406.

Crichton, R., Engelman, D. M., Hass, J., Koch, M. H. J., Moore, P. B., Parfait, R. & Stuhrmann, H. B. (1977). *A contrast variation study of specifically deuterated ribosomal subunits. Proc. Natl Acad. Sci. USA*, **74**, 5547–5550.

Debye, P. (1915). *Zerstreuung von Röntgenstrahlen. Ann. Phys. (Leipzig)*, **46**, 809–823.

Debye, P. & Bueche, A. M. (1949). *Scattering by an inhomogeneous solid. J. Appl. Phys.* **20**, 518–525.

Debye, P. & Pirenne, M. H. (1938). *Über die Fourieranalyse von interferometrischen Messungen an freien Molekülen. Ann. Phys. (Leipzig)*, **33**, 617–629.

Dessen, P., Blanquet, S., Zaccai, G. & Jacrot, B. (1978). *Antico-operative binding of initiator transfer RNA Met to methionyl-transfer RNA synthetase from E. coli: neutron scattering studies. J. Mol. Biol.* **126**, 293–313.

Engelman, D. M. & Moore, P. B. (1972). *A new method for the determination of biological quarternary structure by neutron scattering. Proc. Natl Acad. Sci. USA*, **69**, 1997–1999.

Engelman, D. M. & Zaccai, G. (1980). *Bacteriorhodopsin is an inside-out protein. Proc. Natl Acad. Sci. USA*, **77**, 5894–5898.

Fujiwara, S. & Mendelson, R. A. (1996). *In situ shape and distance measurements in neutron scattering and diffraction. Basic Life Sci.* **64**, 385–395.

Glatter, O. & Kratky, O. (1982). *Small Angle X-ray Scattering*. London: Academic Press.

Guinier, A. (1939). *Diffraction of X-rays of very small angles – application to the study of ultramicroscopic phenomena. Ann. Phys. (Paris)*, **12**, 161–237.

Guinier, A. (1955). *Small Angle Scattering of X-rays*. New York: John Wiley and Sons.

Guinier, A. (1962). *Small angle X-ray scattering*. In *International Tables for X-ray Crystallography*, Vol. III. *Physical and Chemical Tables*, pp. 324–329. Birmingham: Kynoch Press. (Present distributor Kluwer Academic Publishers, Dordrecht).

Hoppe, W. (1972). *New X-ray method for the determination of the quaternary structure of protein complexes. Isr. J. Chem.* **10**, 321–333.

Hunt, J. F., McCrea, P. D., Zaccai, G. & Engelman, D. M. (1997). *Assessment of the aggregation state of integral membrane proteins in reconstituted phospholipid vesicles using small angle neutron scattering. J. Mol. Biol.* **273**, 1004–1019.

Jacrot, B. & Zaccai, G. (1981). *Determination of molecular weight by neutron scattering. Biopolymers*, **20**, 2413–2426.

Jeanteur, D., Pattus, F. & Timmins, P. A. (1994). *Membrane-bound form of the pore-forming domain of colicin A. A neutron scattering study. J. Mol. Biol.* **235**, 898–907.

Junemann, R., Burkhardt, N., Wadzack, J., Schmitt, M., Willumeit, R., Stuhrmann, H. B. & Nierhaus, K. H. (1998). *Small angle scattering in ribosomal structure research: localization of the messenger RNA within ribosomal elongation states. J. Biol. Chem.* **379**, 807–818.

Kratky, O. & Worthmann, W. (1947). *Determination of the configuration of dissolved organic molecules by interferometric measurements with X-rays. Monatsch. Chem.* **76**, 263–281.

Lehmann, M. S. & Zaccai, G. (1984). *Neutron small-angle scattering studies of ribonuclease in mixed aqueous solutions and determination of the preferentially bound water. Biochemistry*, **23**, 1939–1942.

May, R. P., Nowotny, V., Nowotny, P., Voss, H. & Nierhaus, K. H. (1992). *Inter-protein distances within the large subunit from Escherichia coli ribosomes. EMBO J.* **11**, 373–378.

Moore, P. B. (1980). *Small-angle scattering. Information content and error analysis. J. Appl. Cryst.* **13**, 168–175.

Moore, P. B. & Engelman, D. M. (1976). *The production of deuterated E. coli. Brookhaven Symp. Biol.* **27**, V12–V23.

Moore, P. B., Engelman, D. M. & Schoenborn, B. P. (1974). *Asymmetry in the 50S ribosomal subunit of Escherichia coli. Proc. Natl Acad. Sci. USA*, **71**, 172–176.

Moore, P. B., Langer, J. A. & Engelman, D. M. (1978). *The measurement of the locations and radii of gyration of proteins in the 30S ribosomal unit of E. coli by neutron scattering. J. Appl. Cryst.* **11**, 479–482.

Moore, P. B. & Weinstein, E. (1979). *On the estimation of the locations of subunits within macromolecular aggregates from neutron interference data. J. Appl. Cryst.* **12**, 321–326.

Nierhaus, K. H., Lietzke, R., May, R. P., Nowotny, V., Schulze, H., Simpson, K., Wurmbach, P. & Stuhrmann, H. B. (1983). *Shape determinations of ribosomal proteins in situ. Proc. Natl Acad. Sci. USA*, **80**, 2889–2893.

Nierhaus, K. H., Wadzack, J., Burkhardt, N., Junemann, R., Meerwinck, W., Willumeit, R. & Stuhrmann, H. B. (1998). *Structure of the elongating ribosome: arrangement of the two tRNAs before and after translocation. Proc. Natl Acad. Sci. USA*, **95**, 945–950.

Nowotny, P., Ruhl, M., Nowotny, V., May, R. P., Burkhardt, N., Voss, H. & Nierhaus, K. H. (1994). *Direct shape determination of ribosomal proteins in solution and within the ribosome by means of neutron scattering. Biophys. Chem.* **53**, 115–122.

Olah, G. A., Rokop, S. E., Wang, C. L., Blechner, S. L. & Trewhella, J. (1994). *Troponin I encompasses an extended troponin C in the Ca(2+)-bound complex: a small-angle X-ray and neutron scattering study. Biochemistry*, **33**, 8233–8239.

Porod, G. (1951). *The X-ray small-angle scattering of close-packed colloid systems. I. Kolloid-Z.* **124**, 83–144.

Porod, G. (1952). *The X-ray small-angle scattering of close-packed colloid systems. II. Kolloid-Z.* **125**, 51–57.

Serdyuk, I. N., Grenader, A. K. & Zaccai, G. (1979). *Study of the internal structure of E. coli ribosomes by neutron and X-ray scattering. J. Mol. Biol.* **135**, 691–707.

Serdyuk, I. N., Pavlov, M. Yu., Rublevskaya, I. N., Zaccai, G. & Leberman, R. (1994). *Triple isotopic substitution method in small angle scattering. Application to the study of the ternary complex EF-Tu·GTp·aminoacyl tRNA. Biophys. Chem.* **53**, 123–130.

Stöckel, P., May, R., Strell, I., Cejka, Z., Hoppe, W., Heumann, H., Zillig, W. & Crespi, H. (1980*a*). *The core subunit structure in RNA polymerase holoenzyme determined by neutron small-angle scattering. Eur. J. Biochem.* **112**, 411–417.

Stöckel, P., May, R., Strell, I., Cejka, Z., Hoppe, W., Heumann, H., Zillig, W. & Crespi, H. L. (1980*b*). *The subunit positions within RNA polymerase holoenzyme determined by triangulation of centre-to-centre distances. Eur. J. Biochem.* **112**, 419–423.

Stöckel, P., May, R., Strell, I., Cejka, Z., Hoppe, W., Heumann, H., Zillig, W., Crespi, H. L., Katz, J. J. & Ibel, K. (1979). *Determination of intersubunit distances and subunit shape parameters in DNA-dependent RNA polymerase by neutron small-angle scattering. J. Appl. Cryst.* **12**, 176–185.

Stone, D. B., Timmins, P. A., Schneider, D. K., Krylova, I., Ramos, C. H., Reinach, F. C. & Mendelson, R. A. (1998). *The effect of regulatory* Ca^{2+} *on the in situ structures of troponin C and troponin I: a neutron scattering study. J. Mol. Biol.* **281**, 689–704.

Stuhrmann, H. B. (1976). *Small-angle scattering of proteins in solution. Brookhaven Symp. Biol.* **27**, IV3–IV19.

Stuhrmann, H. B., Haas, J., Ibel, K., Koch, M. H. & Crichton, R. R. (1976). *Low angle neutron scattering of ferritin studied by contrast variation*. *J. Mol. Biol.* **100**, 399–413.

Stuhrmann, H. B. & Nierhaus, K. H. (1996). *The determination of the in situ structure by nuclear spin contrast variation*. *Basic Life Sci.* **64**, 397–413.

Stuhrmann, H. B., Tardieu, A., Mateu, L., Sardet, C., Luzzati, V., Aggerbeck, L. & Scanu, A. M. (1975). *Neutron scattering study of human serum low density lipoprotein*. *Proc. Natl Acad. Sci. USA*, **72**, 2270–2273.

Svergun, D. I. (1994). *Solution scattering from biopolymers: advanced contrast-variation data analysis*. *Acta Cryst.* A**50**, 391–402.

Svergun, D. I., Barberato, C., Koch, M. H., Fetler, L. & Vachette, P. (1997). *Large differences are observed between the crystal and solution quaternary structures of allosteric aspartate transcarbamylase in the R state*. *Proteins*, **27**, 110–117.

Svergun, D. I., Burkhardt, N., Pedersen, J. S., Koch, M. H., Volkov, V. V., Kozin, M. B., Meerwink, W., Stuhrmann, H. B., Diedrich, G. & Nierhaus, K. H. (1997). *Solution scattering structural analysis of the 70S Escherichia coli ribosome by contrast variation. II. A model of the ribosome and its RNA at 3.5 nm resolution*. *J. Mol. Biol.* **271**, 602–618.

Svergun, D. I., Koch, M. H., Pedersen, J. S. & Serdyuk, I. N. (1996). *Structural model of the 50S subunit of E. coli ribosomes from solution scattering*. *Basic Life Sci.* **64**, 149–174.

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Protein hydration in solution: experimental observation by X-ray and neutron scattering*. *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.

Timmins, P. A., Hauk, J., Wacker, T. & Welte, W. (1991). *The influence of heptane-1,2,3-triol on the size and shape of LDAO micelles. Implications for the crystallisation of membrane proteins*. *FEBS Lett.* **280**, 115–120.

Uberbacher, E. C., Mardian, J. K., Rossi, R. M., Olins, D. E. & Bunick, G. J. (1982). *Neutron scattering studies and modeling of high mobility group 14 core nucleosome complex*. *Proc. Natl Acad. Sci. USA*, **79**, 5258–5262.

Vainshtein, B. K., Sosfenov, N. I. & Feigin, L. A. (1970). *X-ray method for determining the gaps between heavy atoms in macromolecules in solution and its use for studying gramicidin derivatives. Dokl. Akad. Nauk SSSR*, **190**, 574–577.

Vanatalu, K., Paalme, T., Vilu, R., Burkhardt, N., Junemann, R., May, R., Ruhl, M., Wadzack, J. & Nierhaus, K. H. (1993). *Large-scale preparation of fully deuterated cell components. Ribosomes from Escherichia coli with high biological activity*. *Eur. J. Biochem.* **216**, 315–321.

Zaccai, G. & Jacrot, B. (1983). *Small angle neutron scattering. Annu. Rev. Biophys. Bioeng.* **12**, 139–157.

International Tables for Crystallography (2012). Vol. F, Chapter 19.5, pp. 583–592.

Chapter 19.5. Fibre diffraction

R. Chandrasekaran and G. Stubbs

19.5.1. Introduction

Many biopolymers are long helical structures and have a natural tendency to form fibres. This tendency severely impedes the growth of single crystals from these polymers, and even if crystals can be grown, the molecular interactions in the crystals rarely correspond to the biologically significant interactions in the fibres. Conventional macromolecular crystallography is therefore often not applicable to these systems. Fibre diffraction, however, is a powerful technique for determining the structural details of such polymers. It has been used to study a wide variety of biopolymers, ranging from simple polypeptides, polynucleotides and polysaccharides to complex filamentous viruses and cytoskeletal filaments.

Fibres can have relatively high degrees of order, although falling short of true three-dimensional crystallinity. The key difference between fibres and crystals, however, is that in fibres the fundamental structural aggregates, although parallel to each other, are randomly oriented about the fibre axis. Consequently, the diffraction pattern is cylindrically averaged. This cylindrical averaging is the defining characteristic of fibre diffraction.

On the basis of this definition, fibre diffraction may also be considered to include diffraction from many biological membrane specimens, and much of fibre-diffraction theory also applies to membrane diffraction. In general, however, the diffracting units in fibres have helical symmetry, whereas those of membranes do not.

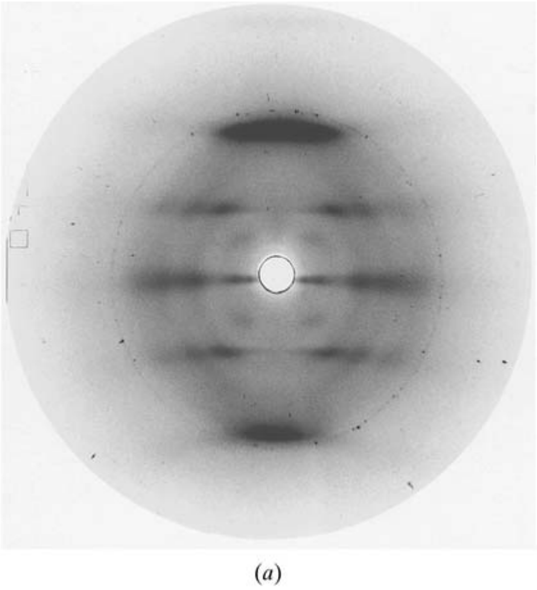
In addition to the loss of information due to cylindrical averaging, fibre-diffraction patterns reflect a generally limited degree of order and rarely extend beyond 3 Å resolution. Consequently, the number of data obtainable from a fibre is considerably less than that from a single crystal having a similar size of asymmetric unit. The use of stereochemical information to supplement the diffraction data is therefore essential. For polymers with small asymmetric units, such as polynucleotides, structural chemical information can be used to construct models consistent with the helical parameters and molecular dimensions obtained from the diffraction data. For the larger asymmetric units found in aggregates, such as viruses, initial models must be constructed in other ways. However, in all cases the combination of diffraction data and stereochemistry can be used to refine both molecular structures and packing parameters. Refinement in this way is very similar to that used in macromolecular crystallography, but because of the limited number of experimental data, stereochemical restraints are particularly important in fibre diffraction. As in crystallography, difference-electron-density maps are used in conjunction with refinement to identify missing portions and determine the correctness of the models and, in favourable cases, to locate ions and solvent molecules associated with the polymers.

19.5.2. Types of fibres

Fibres fall into essentially two classes with respect to the degree of ordering of the polymer chains. Within each class, there are varying degrees of disorder; furthermore, many fibres exhibit properties intermediate between those of the two ideal classes.

In noncrystalline fibres, the polymers are parallel to each other, but their positions and orientations are otherwise uncorrelated. Diffraction patterns from these fibres are confined to layer lines (Fig. 19.5.2.1*a*) because of the repeating nature of the polymer helix, but are otherwise continuous and correspond to the cylindrical average of the Fourier transform of a single particle.

In polycrystalline fibres, the polymers form fully ordered microcrystallites, and each fibre consists of many such microcrystallites, randomly oriented about the fibre axis. In diffraction patterns from polycrystalline fibres, the layer lines are sampled to form discrete reflections (Fig. 19.5.2.1*b*); the diffraction pattern is


(*a*)

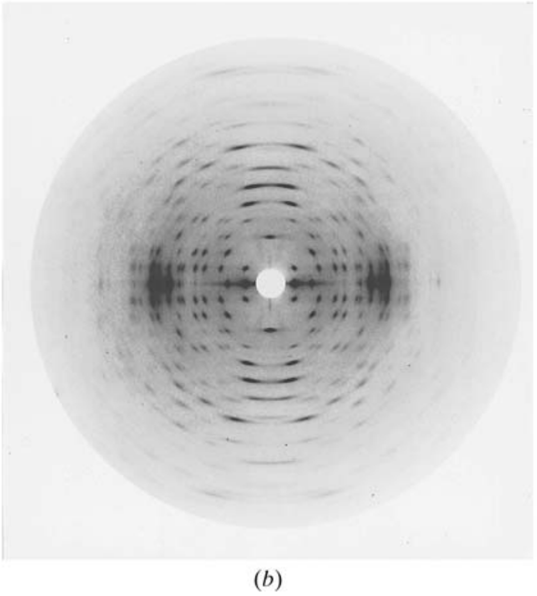

(*b*)

Figure 19.5.2.1
X-ray diffraction patterns showing (*a*) continuous intensity on layer lines from an oriented nucleic acid fibre and (*b*) Bragg reflections from an oriented and polycrystalline polysaccharide fibre.

Copyright © 2012 International Union of Crystallography

the cylindrical average of a single-crystal diffraction pattern and is, in fact, equivalent to the diffraction pattern that would be obtained from a rotating single crystal.

Polycrystalline fibres may be disordered in various ways. For example, the helical polymers may be subject to rotational or translational disorder, and this disorder may be partial (a small number of alternative packings for each particle) or complete (for example, completely random rotational particle orientations). Rotational disorder may be coupled to translational disorder (screw disorder). The resulting diffraction patterns may contain both discrete reflections and continuous diffraction along layer lines; depending upon the type of disorder, the discrete reflections may be confined to the equator (layer line zero) or the low-resolution part of the pattern, or they may be dispersed throughout the pattern. Variations in diffraction effects due to different types of disorder have been discussed by Arnott (1980) and Stroud & Millane (1995).

Fibres are also subject to orientational disorder. The polymer helices in noncrystalline fibres and the microcrystallites in crystalline fibres are not perfectly aligned to the fibre axis; the deviation from parallelism is called the disorientation of the fibre. Disorientation causes the reflections from crystalline fibres and the diffracted intensity from noncrystalline fibres to be spread into arcs (Debye–Scherrer arcs).

19.5.3. Diffraction by helical molecules

19.5.3.1. Fibre diffraction patterns

As noted above, the diffraction pattern from a fibre is confined to layer lines because of the repeating nature of the polymer helix. The layer lines in reciprocal space are perpendicular to the fibre axis in real space. The layer line passing through the origin in reciprocal space is called the equator or zero layer line. The line in a diffraction pattern normal to the equator and passing through the origin is called the meridian. If the fibre axis is perpendicular to the incident X-ray beam, the recorded diffraction pattern is symmetric about both the equator and the meridian. If the fibre is not normal to the incident beam, the pattern is symmetric only about the meridian.

19.5.3.2. Helical symmetry

It is convenient to use cylindrical coordinates to describe helical molecules. In real space we use coordinates (r, φ, z); in reciprocal space (R, ψ, Z). By convention, the z axis is the helix axis and the line $(\varphi = 0, z = 0)$ corresponds to the x axis in Cartesian coordinates. The repeat distance along the z axis is c. Within this distance, there are u repeating units in t turns of the helix. If the coordinates of a point in the first repeating unit are (r, φ, z), then applying the helical symmetry gives the coordinates of the corresponding point in the $(k + 1)$th repeating unit as $(r, \varphi + 2\pi kt/u, z + kc/u)$.

19.5.3.3. Structure factors

Cochran *et al.* (1952) showed that the structure factor on layer line l of a helix made up of repeating subunits is

$$\mathbf{F}(R,\psi,Z) = \sum_j \sum_n f_j J_n(2\pi R r_j) \exp\{i\,[n(\psi + \pi/2) - n\varphi_j + 2\pi l z_j/c]\}. \tag{19.5.3.1}$$

Diffraction occurs only for $Z = l/c$. r_j, φ_j and z_j are the real-space coordinates of atom j in the repeating unit of the helix; f_j is the atomic scattering factor of that atom. J_n is the Bessel function of the first kind of order n. The summation over n includes only those values of n that satisfy the selection rule

$$l = tn + um, \tag{19.5.3.2}$$

where m is any integer. In practice, the summation may be limited to values of $|n|$ less than $2\pi r_{\max} R + 2$, where $r_{\max}$ is the radius of the outermost atom in the polymer, because the value of a Bessel function $J_n(x)$ is negligible for n greater than about $x + 2$. For low-order Bessel functions or applications requiring greater accuracy, slight variations of this limitation are used.

The structure factor **F** is a complex number with an amplitude and phase, and is fully equivalent to that derived using the trigonometric functions in crystallography. The expression for intensity $I = \mathbf{F}\mathbf{F}^* = |\mathbf{F}|^2$ holds good.

Equation (19.5.3.1) can be rewritten (Klug *et al.*, 1958) as

$$\mathbf{F}(R, \psi, l/c) = \sum_n \mathbf{G}_{n,l}(R) \exp[in(\psi + \pi/2)], \tag{19.5.3.3}$$

where the Fourier–Bessel structure factor $\mathbf{G}_{n,l}(R)$ is independent of ψ and is given by

$$\mathbf{G}_{n,l}(R) = \sum_j f_j J_n(2\pi R r_j) \exp[i(-n\varphi_j + 2\pi l z_j/c)]. \tag{19.5.3.4}$$

$J_n(0)$ is 1 when $n = 0$ and 0 otherwise. For this reason, the structure factors on the meridian $(R = 0)$ are nonzero only on layer lines for which l is an integral multiple of u. Hence, a visual inspection of the diffraction pattern often helps to determine u.

19.5.3.4. Fourier–Bessel syntheses

Electron densities may be calculated for crystalline fibres, as they are in crystallography, using Fourier syntheses with coefficients determined for the crystalline reflections.

For noncrystalline fibres, it is more convenient to use Fourier–Bessel syntheses: the electron density ρ at point (r, φ, z) is

$$\rho(r, \varphi, z) = (1/c) \sum_l \sum_n \mathbf{g}_{n,l}(r) \exp[i(n\varphi - 2\pi l z/c)], \tag{19.5.3.5}$$

where

$$\mathbf{g}_{n,l}(r) = \int_0^\infty \mathbf{G}_{n,l}(R) J_n(2\pi R r) 2\pi R \,\mathrm{d}R. \tag{19.5.3.6}$$

19.5.3.5. Diffracted intensities: noncrystalline fibres

The intensity in the diffraction pattern of a noncrystalline fibre is the cylindrical average of the square of the Fourier transform (Franklin & Klug, 1955):

$$\begin{aligned} I(R, l) &= \langle |\mathbf{F}(R, \psi, l/c)|^2 \rangle_\psi \\ &= \sum_n \mathbf{G}_{n,l}(R)\mathbf{G}^*_{n,l}(R) \\ &= \sum_n |\mathbf{G}_{n,l}(R)|^2. \end{aligned} \tag{19.5.3.7}$$

The intensity varies continuously as a function of R along each layer line (Fig. 19.5.2.1*a*).

19.5.3.6. Diffracted intensities: polycrystalline fibres

The intensity in the diffraction pattern of a polycrystalline fibre consists of Bragg reflections on layer lines (Fig. 19.5.2.1*b*). On each layer line, owing to the lattice sampling that arises from the lateral organization of the polymers, intensities are observed at discrete R values defined by the reciprocal-lattice points. In the case of monoclinic (with c as the unique axis), orthorhombic and hexagonal systems, the reflection positions are determined by equations (19.5.3.8), (19.5.3.9) and (19.5.3.10), respectively.

$$R_{hk}^2 = h^2a^{*2} + k^2b^{*2} + 2hka^*b^*\cos\gamma^*, \qquad (19.5.3.8)$$

$$R_{hk}^2 = h^2a^{*2} + k^2b^{*2}, \qquad (19.5.3.9)$$

$$R_{hk}^2 = (h^2 + k^2 + hk)a^{*2}. \qquad (19.5.3.10)$$

Consequently, on each layer line, superposition occurs between reciprocal-lattice points (hkl) and $(\bar{h}\bar{k}l)$ for monoclinic; (hkl), $(\bar{h}kl)$, $(h\bar{k}l)$ and $(\bar{h}\bar{k}l)$ for orthorhombic; and (hkl), $(\bar{h}\bar{k}l)$, (khl), $(\bar{k}\bar{h}l)$, (kil), $(\overline{ki}l)$, (ikl), $(\overline{ik}l)$, (ihl), $(\overline{ih}l)$, (hil) and $(\overline{hi}l)$, where $i = -(h + k)$, for hexagonal systems. Depending upon the unit-cell dimensions, other reflections having the same R value may also be superposed to give a single intensity, and those having R values close to each other may be difficult to resolve. All superposed reflections must be considered individually when calculating such composite intensities.

19.5.4. Fibre preparation

Natural fibre specimens may require only the selection of fibres or regions of fibres in which the polymers are well oriented, but many other fibres must be made in the laboratory, and orientation can often be improved by a wide variety of procedures.

Orientation often requires controlled conditions of relative humidity and temperature during the preparation of the fibre; in many cases, these conditions must be maintained during data collection. In some cases, tension must be applied to the fibres; in an increasing number of cases, magnetic fields have been found to improve orientation. Fibres may be drawn directly from concentrated polymer solutions or made by stretching gels, using weights to stretch strips of polymer films cast on Teflon blocks or applying radial heating while forming polymer films (Arnott, Guss *et al.*, 1974; Chandrasekaran, Radha, Lee & Zhang, 1994). A drop of concentrated polymer solution may simply be dried while suspended between two supports (McDonald *et al.*, 2008). Magnetic fields have dramatically improved the orientation in dried fibres of polymers having significant dipole moments (Torbet, 1987).

Oriented sols, generally enclosed in glass capillaries, are usually made using shearing forces, either by moving the sol in the capillary (Gregory & Holmes, 1965; Kendall & Stubbs, 2006) or by centrifugation (Cohen *et al.*, 1971). Again, magnetic fields can greatly improve orientation, sometimes in combination with centrifugation (Yamashita, Suzuki & Namba, 1998).

Any of these stretching or orienting processes might facilitate the growth of long microcrystallites along the fibre axis. Crystallization in general and lateral organization in particular are achieved primarily by careful choice of solution conditions, including solvent, pH, additives, relative humidity and temperature. In both crystalline and noncrystalline specimens, annealing processes are often important to both crystallization and orientation.

19.5.5. Data collection

Fibre-diffraction data were historically collected using laboratory X-ray sources and photographic film, but synchrotron sources (Shotton *et al.*, 1998) are now generally used for final data collection (Janaswamy & Chandrasekaran, 2005, 2006; Tsuruta & Irving, 2008), taking advantage of reduced exposure time, the potential for time-resolved studies and the fact that many fibres (or the well oriented regions of fibres) are too small for laboratory data collection. Charge-coupled device (CCD) detectors have largely replaced film. Pinhole cameras, mirror-monochromator optics and double-mirror optics are used in different applications. Diffraction by most fibres is inherently weak, and very long repeat distances often require long distances between the specimen and the detector, so fibre cameras are often flushed with helium or kept under vacuum to reduce air scatter. Constant relative humidity is often required, and is achieved by bubbling a helium stream through a saturated salt solution followed by a salt trap or by isolating the sample in equilibrium with a saturated salt solution (McDonald *et al.*, 2008).

The X-ray beam commonly strikes a stationary fibre perpendicular to the fibre axis. Because of the cylindrical averaging of the data, this procedure allows most of the diffraction pattern to be collected in a single exposure. There is, however, a ‘blind region’ around the meridian, where the Ewald sphere does not intersect the diffraction pattern (Fraser *et al.*, 1976). Data in this region are collected by tilting the fibre.

19.5.6. Data processing

19.5.6.1. *Coordinate transformation*

Data must be transformed from detector space into reciprocal space (Fraser *et al.*, 1976). Transformation of coordinates requires determination of the origin of the diffraction pattern in detector space, the fibre tilt angle β, the twist angle (often called in-plane tilt, the inclination of the projection of Z along the beam to the detector coordinate system) and the specimen-to-detector distance. It may also require determination of the detector missetting angles (the deviation of the normal to the detector plane from the beam). All of these parameters can be determined by comparing equivalent reflections in the diffraction pattern.

Most data-processing programs determine the transformation parameters by some form of minimization of the deviation from equivalence in the positions of well resolved equivalent reflections. The tilt was traditionally determined by comparing the apparent Z values of equivalent reflections, but the apparent value of R for near-meridional reflections is much more sensitive to tilt. The minimization set should therefore include some near-meridional reflections if the tilt value is to be determined accurately. The helical repeat distance and, for polycrystalline fibres, the unit-cell parameters must also be determined at this time, but helical repeat distance and specimen-to-detector distance are so highly correlated that it is not often practical to refine both. Data may be visualized and in many cases transformed and processed using a number of software applications, including the programs *WCEN* (Bian *et al.*, 2006), *FibreFix* (Rajkumar *et al.*, 2005) and *FIT2D* (http://www.esrf.eu/computing/scientific/FIT2D).

19.5.6.2. *Intensity correction*

Data intensities must be corrected for geometric and polarization effects (Fraser *et al.*, 1976; Millane & Arnott, 1986). The geometric correction has two components: a factor due to the geometry of the intersection between the diffraction pattern in reciprocal space and the sphere of reflection, and a factor due to the angle of incidence of the diffracted beam on the detector. The first factor is analogous to the Lorentz factor in crystallography, which arises because of the time taken for a reflection from a moving sample to pass through the Ewald sphere. The geometric correction can be applied to each data point as a single correction (Fraser *et al.*, 1976). For both crystalline and noncrystalline fibres, it has often been convenient to integrate reflections or arcs before applying Lorentz and polarization corrections, or to apply some corrections before integration and some after (Millane &

Arnott, 1986), but this practice is decreasing as computational power increases. The Lorentz correction for crystalline fibres is

$$1/L = 2\pi \sin\theta\,[\cos^2\theta\cos^2\beta - (\cos\sigma - \sin\theta\sin\beta)^2]^{1/2}, \quad (19.5.6.1)$$

where θ is the Bragg angle and $\tan\sigma = R/Z$ (Millane & Arnott, 1986). The polarization correction for an unpolarized X-ray beam is

$$p = (1 + \cos^2 2\theta)/2. \quad (19.5.6.2)$$

For a polarized beam, such as a synchrotron beam, the expression for p is more complicated (Azároff, 1955; Kahn *et al.*, 1982; Bian *et al.*, 2006). Intensities should be divided by Lp. Intensities may also be corrected for nonlinearity of detector response and for absorption by the specimen and by detector components.

19.5.6.3. Background subtraction

The background can be very high in fibre-diffraction data because of long exposure times and scattering from amorphous material. Because of specimen disorientation, fibre-diffraction data often contain large regions where there is no space between layer lines, so local-background-fitting methods are rarely useful. The background may be determined by fitting an analytical function to intensities at points between reflections (Millane & Arnott, 1985; Lorenz & Holmes, 1993), or by fitting a function that includes both signal and background components to the reflection data. This type of profile fitting has been described for individual reflections (Fraser *et al.*, 1976), for data in concentric rings about the centre of the diffraction pattern (Makowski, 1978) and for entire data sets (Yamashita *et al.*, 1995; Ivanova & Makowski, 1998).

19.5.6.4. Integration of crystalline fibre data

The variation of reflection shape in detector space can be determined using a few sharp reflections and taking into account parameters related to crystallite size and disorientation in the specimen (Millane & Arnott, 1986). This allows the integration boundary of a reflection to be determined. Sometimes, the boundary encompasses two or more reflections too close to separate; such reflections are considered to constitute a composite reflection.

19.5.6.5. Integration of continuous data

In diffraction from noncrystalline fibres, intensity is a function of R on each layer line. Angular deconvolution (Makowski, 1978; Namba & Stubbs, 1985; Yamashita *et al.*, 1995) or profile fitting (Millane & Arnott, 1986) corrects for disorientation and overlap between adjacent layer lines and may also incorporate background subtraction. The intensity determined in this way should be corrected for geometric and other effects if this has not been done previously (Section 19.5.6.2; Namba & Stubbs, 1985; Millane & Arnott, 1986).

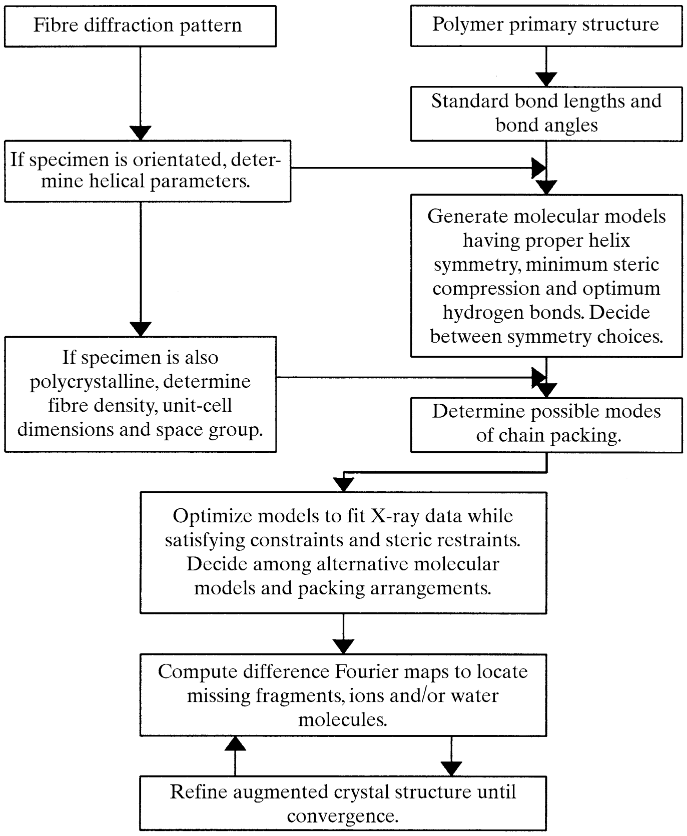

Fibre diffraction pattern
Polymer primary structure
Standard bond lengths and bond angles
If specimen is orientated, determine helical parameters.
Generate molecular models having proper helix symmetry, minimum steric compression and optimum hydrogen bonds. Decide between symmetry choices.
If specimen is also polycrystalline, determine fibre density, unit-cell dimensions and space group.
Determine possible modes of chain packing.
Optimize models to fit X-ray data while satisfying constraints and steric restraints. Decide among alternative molecular models and packing arrangements.
Compute difference Fourier maps to locate missing fragments, ions and/or water molecules.
Refine augmented crystal structure until convergence.

Figure 19.5.7.1
Flow chart of the principal steps in the determination and refinement of fibre structures with small unit cells.

19.5.7. Determination of structures

If the amplitude and phase of each diffracted wave are known, structure determination is, in principle, straightforward (Section 19.5.3.4). In practice, however, the phase problem for fibres is more acute than for single crystals because of the limited resolution of the data, and because the diffracted intensities overlap as a result of disorientation and cylindrical averaging. Patterson methods (MacGillavry & Bruins, 1948; Stubbs, 1987) have sometimes been useful, but the cylindrically averaged Patterson function is usually too complicated for detailed interpretation. Phasing by heavy-atom methods is not practical for polymers with small unit cells because of the difficulties in incorporating heavy atoms into the structures. Structures having small unit cells are instead determined by constructing initial models based on chemical information and the observed helical parameters. Extensions of the isomorphous-replacement method (Namba & Stubbs, 1985) have been useful in determining structures, such as those of helical viruses, in which the unit cells are much larger. In all cases, refinement and evaluation of the model structures are essential. A flow chart of the sequential steps in the determination and refinement of fibre structures with small unit cells is shown in Fig. 19.5.7.1.

19.5.7.1. Initial models: small unit cells

For many biopolymers, especially polypeptides, polynucleotides and polysaccharides, the repeating unit is a monomer or a small oligomer and the unit-cell dimensions are in the range 10 to 50 Å. Such unit cells can accommodate one or more polymer helices, packed in an organized fashion.

An initial model is constructed from the primary structure of the repeating unit, using bond lengths, bond angles and some conformation angles derived from surveys of accurate single-crystal analyses. The model must satisfy the observed helical parameters and have reasonable intra- and inter-chain nonbonded, hydrogen-bonded and polar interactions.

This preliminary model provides an approximate solution to the phase problem and a starting point for refinement. Since there is no assurance that the refined model represents the true structure, however, stereochemically plausible alternatives must be carefully considered, refined and objectively adjudicated.

Alternatives can include both right- and left-handed helices, single helices, and multistranded helices with parallel and antiparallel strands. The next stage involves the packing arrangement in the unit cell. If two or more helices are present, their positions, orientations and relative polarities must be varied in refinement.

19.5.7.2. Refinement: small unit cells

The widely used linked-atom least-squares (*LALS*) technique (Arnott & Wonacott, 1966; Smith & Arnott, 1978) and the variable virtual bond (*PS79*) method (Zugenmaier & Sarko, 1980) were developed for fibre structures. They are similar in principle to the least-squares refinement procedure for crystalline proteins (Hendrickson, 1985), although bond lengths and bond angles are usually kept fixed in the fibre refinements. The function minimized by the *LALS* program is of the form

$$\Omega = \sum_m w_m \Delta F_m^2 + \sum_i e_i \Delta\theta_i^2 + \sum_j k_j \Delta c_j^2 + \sum_n \lambda_n G_n. \qquad (19.5.7.1)$$

The first term on the right-hand side is the weighted sum of the squares of the differences, ΔF_m, between observed and calculated X-ray structure amplitudes of Bragg reflections or continuous diffraction. Either or both types of data can be used as necessary. The weights, w_m, are inversely proportional to the estimated variance of the data. The second term minimizes the differences, $\Delta\theta_i$, between the expected (standard) values of conformation and bond angles and those in the model; the weights, e_i, are based on empirically determined variances. The third term is designed to take care of non-bonded interactions and thus keep the model free from steric compression. It includes the deviations from target values of both intra- and inter-chain hydrogen bonds and the differences between acceptable and calculated non-bonded distances for those contacts that are smaller than the acceptable limiting values. The weights, k_j, are based on the Buckingham energy function for non-bonded contacts and empirical variances for hydrogen bonds. Finally, the fourth term imposes constraints (G_h, with Lagrange multipliers λ_h) for helix connectivity and ring closure, as in a furanose or pyranose, and it vanishes when all such constraints are satisfied. During the refinement, the structure factors are calculated with either the conventional atomic scattering factor f or with a solvent-corrected atomic scattering factor f_w (Fraser *et al.*, 1978; Chandrasekaran & Radha, 1992) given by the function

$$f_w(D) = f(D) - v\sigma_s \exp(-\pi v^{2/3} D^2), \qquad (19.5.7.2)$$

where $D = (R^2 + Z^2)^{1/2}$, σ_s is the electron density of the solvent and v is the excluded volume of the atom. If the van der Waals radius of water is taken as 2 Å, σ_s for water is 0.2984 e Å^{-3}. Equation (19.5.7.2) allows for the solvent contribution to the diffracted intensity and is particularly useful in studying hydrated fibres in which structured and amorphous water can account for up to 50% of the total mass.

19.5.7.3. Data-to-parameter ratio

The total number of data used in this refinement process is $M + I + J$, where M, I and J are, respectively, the number of observations in the first three terms of equation (19.5.7.1). If P is the number of parameters refined and H is the number of independent constraints in the last term, then the number of degrees of freedom of the system is $P - H$. The effective number of data is given by $D = (M + I + J) - (P - H)$. The data-to-parameter ratio (*D/P*), a measure of the dependability of the final results, must be greater than one for meaningful refinement. *D/P* is typically in the range 3 to 11 in the analysis of polynucleotide and polysaccharide structures. This ratio is comparable to those commonly reported for single-crystal structures, confirming that fibre-diffraction analysis of polymers, despite the limited number of X-ray data, can yield reliable results.

19.5.7.4. Initial models: large unit cells

For large macromolecular aggregates, such as viruses and cytoskeletal filaments, initial models cannot usually be devised using the primary structure of the molecule alone. The largely α-helical filamentous bacteriophages form a rare class of exceptions (Makowski *et al.*, 1980). Molecular-replacement methods, in which initial models are constructed from single-crystal structure determinations of the separated components of the aggregate or from known related structures, can be useful, but because of the limited number of data in a fibre pattern such models can sometimes be difficult to refine.

Multi-dimensional isomorphous replacement (MDIR), an extension of the isomorphous-replacement method of protein crystallography, has been useful in studying helical viruses (Stubbs & Diamond, 1975; Namba & Stubbs, 1985). The dimensions are the real and imaginary parts of the various overlapping structure factors at a given point in the diffraction pattern. Information about both the phases of the structure factors and the relative magnitudes of the overlapping structure factors is obtained from heavy-atom derivatives of the virus; at least twice as many heavy-atom derivatives as the number of significant **G** terms in equation (19.5.3.7) are required. If the structure of a related aggregate is known, MDIR can be combined with molecular replacement (Namba & Stubbs, 1987*a*; Wang & Stubbs, 1994); in this case, fewer derivatives are required.

Layer-line splitting (Franklin & Klug, 1955) arises when the helical symmetry of the scattering particles is close to, but not exactly, integral. For example, tobacco mosaic virus (TMV) has 49.02 subunits in three turns of the viral helix. In this case, the **G** terms in each layer line do not fall at exactly the same Z values in the diffraction pattern. The resulting shifts in the positions of the layer lines can be measured for the native aggregate and, in favourable cases, for heavy-atom derivatives, and used to provide additional phase information (Stubbs & Makowski, 1982). Information from electron microscopy (Beese *et al.*, 1987) and neutron scattering (Nambudripad *et al.*, 1991) has also been used.

19.5.7.5. Refinement: large unit cells

Refinement of fibre structures having large unit cells has many parallels to refinement in protein crystallography. Refinement in real space, especially the solvent-flattening approach, has been widely used to improve electron-density maps and is particularly valuable in structure determination of noncrystalline fibres. Since helical aggregates have finite radii, **g** terms [equation (19.5.3.6)] can be set to zero outside a maximum radius and back-transformed to obtain refined estimates of the phases of the **G** terms. More detailed solvent-flattening algorithms can also be used (Namba & Stubbs, 1985).

Molecular models can be refined by methods conceptually related to those of *LALS*. The principal difference is that bond lengths and angles are not kept fixed, but are restrained to remain close to standard values. The restrained least-squares method (Hendrickson, 1985), widely used in protein crystallography, has been adapted (Stubbs *et al.*, 1986) for fibre diffraction and used to refine a number of filamentous virus structures (Namba *et al.*, 1989; Nambudripad *et al.*, 1991). Although effective, the radius of convergence of this method is less than desired, probably because of the limited number of data available from fibre diffraction (Wang & Stubbs, 1993).

Molecular-dynamics methods have been used to increase the radius of convergence of refinement (Wang & Stubbs, 1993). The program *X-PLOR* (Brünger *et al.*, 1987) has been adapted for fibre diffraction and can handle data from both crystalline and noncrystalline fibres. A potential-energy function of the form

$$\Omega = E + S \sum_{l} \sum_{i} w_{li} \{[I_o(R_i)]^{1/2} - [I_c(R_i)]^{1/2}\}^2 \qquad (19.5.7.3)$$

is minimized. The first term, E, is an empirical energy function that accounts for distortions in bond lengths, bond angles and conformation angles, and for non-bonded, electrostatic and hydrogen-bonding interactions. The second term accounts for the differences between the observed and calculated X-ray intensities at specific values of R_i on every layer line l; w_{li} is the weight for each observation and S is a normalizing factor. In the most effective use of this method, simulated annealing, the process of heating the structure to a temperature of 3000 to 4000 K is simulated, and then the structure is cooled ('annealed') in small increments. At high temperatures, energy barriers between the starting model and structures of lower potential can be overcome; in this way, the radius of convergence of the refinement is increased.

19.5.7.6. Difference Fourier methods

As in crystallography, difference maps are used during refinement to correct errors and to identify missing fragments of the model and, in the final stages of refinement, to identify solvent molecules and associated ions.

In crystalline fibre diffraction, the most common difference maps use calculated phases with amplitudes of either $F_o - F_c$ or $2F_o - F_c$. In both cases, weighting the coefficients on the basis of the observed and calculated structure amplitudes has been used to minimize the root-mean-square error in the electron-density maps. Reflections superposed by cylindrical averaging do, however, present problems. One solution is to divide the observed intensity equally among the superposed reflections. This is a reasonable approach in the initial stages of structure analysis, when the reliability of the model is uncertain, and has the advantage of minimizing bias toward the model. Alternatively, the observed intensity may be split in the same ratio as the calculated intensity. This approach, although biased, is more effective for locating solvent molecules and ions in an otherwise well determined structure. Difference Fourier maps have played a significant role in determining the molecular structures and packing arrangements in unit cells mediated by water molecules and cations of several polynucleotide (Chandrasekaran *et al.*, 1995, 1997) and polysaccharide helices (Winter *et al.*, 1975; Chandrasekaran *et al.*, 1988, 1998; Chandrasekaran, Radha & Lee, 1994; Janaswamy & Chandrasekaran, 2001, 2002, 2008).

In noncrystalline fibre diffraction, the superposition of intensities due to cylindrical averaging is more serious and must be taken into account. Namba & Stubbs (1987*b*) have shown that the coefficients yielding the most accurate electron-density maps of the full structure have amplitudes of $NG_o - (N-1)G_c$, where N is the number of significant terms in equation (19.5.3.7) (the number of superposed intensities), and the observed intensity is divided in the ratio of the calculated intensity. For filamentous viruses at moderate resolution, N is typically in the range four to six. As in crystallography and crystalline fibre diffraction, maps calculated from amplitudes of $F_o - F_c$ have low noise levels and are most useful for checking the accuracy of final models and for locating solvent molecules.

19.5.7.7. Evaluation

As in crystallography, fibre structures are evaluated by statistical measures, such as R values, and by the examination of difference maps. Fibre-diffraction R values are inherently lower than those expected in crystallography, particularly when large numbers of intensities have been superposed by cylindrical averaging (Stubbs, 1989). The largest likely R value for noncrystalline TMV at 3 Å resolution is about 0.31 and for polycrystalline DNA at 3 Å resolution it is about 0.41, both significantly less than the value of 0.59 to be expected from noncentric single-crystal analyses (Millane, 1989). Free R values, widely used in crystallography, have been used to a limited extent in fibre diffraction (Welsh *et al.*, 2000), sometimes with modifications to deal with the effect of overlapping Bessel orders (Welsh *et al.*, 1998). Some authors prefer to use correlation coefficients to evaluate models.

Comparison of R values alone is not necessarily a reliable way to discriminate between competing models. Such discrimination is often required for structures with small unit cells, for which alternative models are routinely refined (Sections 19.5.7.1 and 19.5.7.2). The relative merits of any pair of competing models can be assessed on the basis of several types of statistics (Arnott, 1980) using Hamilton's significance test (Hamilton, 1965), which considers not only residuals but also numbers of degrees of freedom (Section 19.5.7.3). Such a test is essential. There are many examples in the literature where R values have been lowered by the simple process of increasing the number of degrees of freedom; a decreased R value obtained in this way may or may not have any significance.

Difference Fourier maps have been used to evaluate crystalline fibre diffraction analyses for many years, for example, to reject the controversial Hoogsteen base pairing in double-stranded DNA (Arnott *et al.*, 1965), and later to discriminate between 10- and 11-fold double helices of RNA (Arnott *et al.*, 1967). Difference maps have been essential in the refinement of fibre structures with large unit cells (Namba *et al.*, 1989; Wang & Stubbs, 1994), both to identify errors in early models and to confirm that the final structures contained no major errors or omissions.

19.5.8. Structures determined by X-ray fibre diffraction

The α-helix of several synthetic polypeptides (Pauling & Corey, 1951), the double helix of DNA (Franklin & Gosling, 1953*a*,*b*; Watson & Crick, 1953), the ribbon structure of cellulose (Meyer & Misch, 1937) and the low-resolution structure of tobacco mosaic virus (Barrett *et al.*, 1971) were early examples of structures determined by fibre diffraction. Early workers also examined a number of fibrous proteins (Bailey *et al.*, 1943). In the past 50 years, developments in theory and practice and the availability of fast computers have made it possible to determine and refine about 200 biological polymer structures of varying complexities. The largest repeating units in polypeptides, polynucleotides and polysaccharides solved to date correspond to a tripeptide, a tetranucleotide and a heptasaccharide, respectively.

19.5.8.1. Polypeptides

The structural details of the α-helix and β-sheet, the principal secondary-structure elements of proteins, have emerged from the analysis of synthetic polypeptides (Pauling & Corey, 1951, 1953).

Analysis of noncrystalline fibre-diffraction patterns led to the triple-helical coiled-coil model of collagen (Ramachandran & Kartha, 1955; Rich & Crick, 1955). There have been many recent studies on amyloid fibrils, associated with a large number of diseases including Alzheimer's and Parkinson's diseases, and including peptide fragments of prions, infectious amyloids associated with diseases such as bovine spongiform encephalopathy ('mad cow disease') (Inouye *et al.*, 1993, 2000; Malinchik *et al.*, 1998; Kishimoto *et al.*, 2004; Makin & Serpell, 2005).

19.5.8.2. Polynucleotides

The molecular structures of a series of DNA and RNA helices have been determined and refined using data from polycrystalline fibres (Arnott *et al.*, 1969; Chandrasekaran & Arnott, 1989). These include the canonical A, B and C forms of DNA, corresponding, respectively, to 11-, 10- and 9.3-fold right-handed antiparallel Watson–Crick base-paired helices. Structural differences between the three have been attributed to changes in furanose puckerings and helical parameters: the A form has C3-*endo*, but B and C have C2-*endo* or analogous C3-*exo* puckers. All RNA duplexes are members of the A family. Later important structures included the sixfold single helix of poly (C) (Arnott *et al.*, 1976), a compact eightfold double helix for poly d(AT) and poly d(IC) (Arnott *et al.*, 1983), and the left-handed Z-DNA for poly d(GC) (Arnott *et al.*, 1980). Difference Fourier syntheses were instrumental in locating a spine of water molecules in the minor groove and a series of sodium ions and water molecules that bridge the phosphate groups of adjacent DNA molecules in the tenfold helices of poly (dA)·poly (dT) (Chandrasekaran *et al.*, 1995), poly (dA)·poly (dU) and poly d(AI)·poly d(CT) (Chandrasekaran *et al.*, 1997). Neutron fibre diffraction has been particularly useful for locating water molecules in polynucleotides (Shotton *et al.*, 1997, 1998). Data from noncrystalline fibres have been used to determine, among others, the structures of DNA·RNA hybrid duplexes (Arnott *et al.*, 1986), a DNA triple-stranded helix (Chandrasekaran *et al.*, 2000*a*) and two RNA triple-stranded helices (Chandrasekaran *et al.*, 2000*b*,*c*).

19.5.8.3. Polysaccharides

Among the three-dimensional structures determined for industrially useful and biologically important polysaccharides are the gel-forming calcium *i*-carrageenan (Arnott, Scott *et al.*, 1974; Janaswamy & Chandrasekaran, 2001, 2002, 2008), sodium pectate (Walkinshaw & Arnott, 1981), gellan (Chandrasekaran *et al.*, 1988) and welan (Chandrasekaran, Radha & Lee, 1994), and a series of distinct helical forms of the glycosaminoglycan hyaluronan (Arnott & Mitra, 1984). The conformations of these molecules are delicately controlled by ions, such as sodium, potassium and calcium. The repeating units range from a simple monosaccharide to a branched pentasaccharide.

Specific interactions among the polysaccharides and their associated small molecules can be correlated with their observed properties. A number of neutral polysaccharides, such as cellulose, chitin and mannan, are twofold ribbon-like helices, which aggregate and are hence water insoluble. The *A* and *B* forms of amylose, the main constituents of starch granules, are sixfold left-handed parallel double helices. Derivatization of amylose leads to the formation of single helices (Chandrasekaran, 1997). The water-soluble galactomannan derives its high viscosity in aqueous solution from intermolecular side-chain interactions (Chandrasekaran *et al.*, 1998). Gels of calcium iota-carrageenan are stronger owing to more intimate sulfate-group interactions between double helices than in the sodium form (Janaswamy & Chandrasekaran, 2001, 2002, 2008).

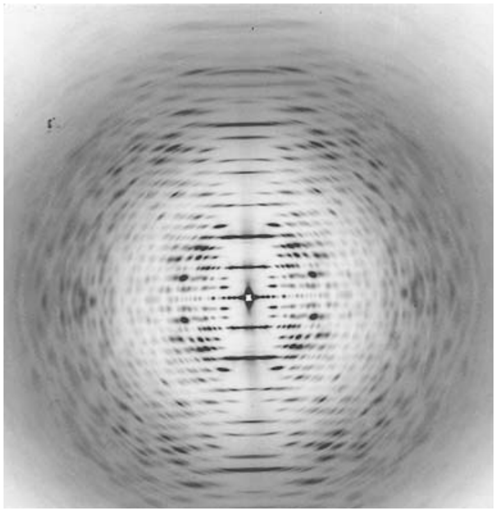


Figure 19.5.8.1
X-ray diffraction pattern from an oriented sol of the U2 strain of tobacco mosaic virus.

19.5.8.4. Helical viruses and bacteriophages

The largest repeating units in structures determined by fibre diffraction are those of several members of the tobamovirus family, including tobacco mosaic virus (Namba *et al.*, 1989; Kendall *et al.*, 2007), cucumber green mottle mosaic virus (Wang & Stubbs, 1994) and ribgrass mosaic virus (Wang *et al.*, 1997). These viruses are rod-shaped, 3000 Å long and about 180 Å in diameter. Oriented sols yield exceptionally good diffraction patterns (Fig. 19.5.8.1). The asymmetric unit consists of a protein subunit of approximate molecular weight 18 000 Da and three nucleotides of RNA. The coat proteins are folded like globular proteins and are about 40% α-helical, with small regions of β-sheet. All of the amino acids, all three nucleotides, and in some cases water molecules and calcium ions, are seen in the electron-density maps. The TMV structure was determined by MDIR; the remaining structures were determined by molecular replacement from TMV or by a combination of molecular replacement and isomorphous replacement. All of the structures were refined by restrained least-squares or molecular-dynamics methods to *R* values of less than 0.10 at resolutions between 2.9 and 3.5 Å. Fibre diffraction data from a number of flexible filamentous viruses, including members of the *Flexiviridae* and *Potyviridae* families, have been used in conjunction with data from electron cryomicroscopy to determine symmetries and low-resolution structures (Kendall *et al.*, 2008).

Several filamentous bacteriophage structures, including fd, Pf1 and related strains, have been determined and refined. Filamentous bacteriophages are flexible viruses, about 60 Å in diameter and 10 000 to 20 000 Å in length. Several thousand copies of a coat protein of about 50 residues wrap around a central single-stranded circular DNA. The DNA does not appear to be sufficiently ordered to appear in electron-density maps. The coat-protein molecules have an unusually simple structure, being almost entirely α-helical (Marvin *et al.*, 1974). Model-building approaches have therefore been used, sometimes supplemented by isomorphous replacement (Bryan *et al.*, 1983). Neutron scattering from bacteriophages with selectively deuterated amino-acid residues has also been used to assist model building

(Nambudripad *et al.*, 1991). Both restrained least-squares (Nambudripad *et al.*, 1991) and molecular-dynamics (Gonzalez *et al.*, 1995) refinement methods have been used. The coat protein forms two α-helical layers (Welsh *et al.*, 1998).

19.5.8.5. Other large assemblies

Low-resolution X-ray fibre-diffraction data have been successfully used to model the structural details of a number of complex assemblies. For example, the structure of the F-actin helix at 8 Å resolution has been described by combining the single-crystal structure of the G-actin monomer with fibre-diffraction data (Holmes *et al.*, 1990; Oda *et al.*, 2009). This structure, in turn, has been used to model the muscle thin filament, composed of F-actin monomers and tropomyosin, at about 25 Å resolution, both in the resting and activated states, and hence to understand the movement of tropomyosin in muscle function (Squire *et al.*, 1993). The structure of the microtubule has been determined at 18 Å resolution using information from electron microscopy and fibre diffraction (Beese *et al.*, 1987). A similar but more sophisticated approach was used for bacterial flagellar filaments at 9 Å resolution (Yamashita, Hasegawa *et al.*, 1998); the diffraction patterns obtained from these filaments are of such high quality that prospects for a complete molecular structure are excellent.

References

Arnott, S. (1980). *Twenty years hard labor as a fiber diffractionist*. *Am. Chem. Soc. Symp. Ser.* **141**, 1–30.

Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W. & Ratliff, R. L. (1980). *Left-handed DNA helices*. *Nature (London)*, **283**, 743–745.

Arnott, S., Chandrasekaran, R. & Leslie, A. G. W. (1976). *Structure of the single-stranded polyribonucleotide polycytidylic acid*. *J. Mol. Biol.* **106**, 735–748.

Arnott, S., Chandrasekaran, R., Millane, R. P. & Park, H.-S. (1986). *DNA–RNA hybrid secondary structures*. *J. Mol. Biol.* **188**, 631–640.

Arnott, S., Chandrasekaran, R., Puigjaner, L. C., Walker, J. K., Hall, I. H. & Birdsall, D. L. (1983). *Wrinkled DNA*. *Nucleic Acids Res.* **11**, 1457–1474.

Arnott, S., Dover, S. D. & Wonacott, A. J. (1969). *Least-squares refinement of the crystal and molecular structures of DNA and RNA from X-ray data and standard bond lengths and angles*. *Acta Cryst.* B**25**, 2192–2206.

Arnott, S., Guss, J. M., Hukins, D. W. L., Dea, I. C. M. & Rees, D. A. (1974). *Conformation of keratan sulphate*. *J. Mol. Biol.* **88**, 175–184.

Arnott, S. & Mitra, A. (1984). *X-ray diffraction analysis of glycosaminoglycans*. In *Molecular Biophysics of the Extracellular Matrix*, edited by S. Arnott, D. A. Rees & E. R. Morris, pp. 41–67. New Jersey: Humana Press.

Arnott, S., Scott, W. E., Rees, D. A. & McNab, C. G. A. (1974). *i-Carrageenan: molecular structure and packing of polysaccharide double helices in oriented fibers of divalent cation salts*. *J. Mol. Biol.* **90**, 253–267.

Arnott, S., Wilkins, M. H. F., Fuller, W. & Langridge, R. (1967). *Molecular and crystal structures of double-helical RNA. III. An 11-fold molecular model and comparison of the agreement between the observed and calculated three-dimensional diffraction data for 10- and 11-fold models*. *J. Mol. Biol.* **27**, 535–548.

Arnott, S., Wilkins, M. H. F., Hamilton, L. D. & Langridge, R. (1965). *Fourier synthesis studies of lithium DNA. Part III: Hoogsteen models*. *J. Mol. Biol.* **11**, 391–402.

Arnott, S. & Wonacott, A. J. (1966). *The refinement of the crystal and molecular structures of polymers using X-ray data and stereochemical constraints*. *Polymer*, **7**, 157–166.

Azároff, L. V. (1955). *Polarization correction for crystal-monochromatized X-radiation*. *Acta Cryst.* **8**, 701–704.

Bailey, K., Astbury, W. T. & Rudall, K. M. (1943). *Members of the keratin-myosin group*. *Nature (London)*, **151**, 716–717.

Barrett, A. N., Leigh, J. B., Holmes, K. C., Leberman, R., Sengbusch, P. & Klug, A. (1971). *An electron density map of tobacco mosaic virus at 10 Å resolution*. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 433–448.

Beese, L., Stubbs, G. & Cohen, C. (1987). *Microtubule structure at 18 Å resolution*. *J. Mol. Biol.* **194**, 257–264.

Bian, W., Wang, H., McCullough, I. & Stubbs, G. (2006). *WCEN: a computer program for initial processing of fiber diffraction patterns*. *J. Appl. Cryst.* **39**, 752–756.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R-factor refinement by molecular dynamics*. *Science*, **235**, 458–460.

Bryan, R. K., Bansal, M., Folkhard, W., Nave, C. & Marvin, D. A. (1983). *Maximum-entropy calculation of the electron density at 4 Å resolution of Pf1 filamentous bacteriophage*. *Proc. Natl Acad. Sci. USA*, **80**, 4728–4731.

Chandrasekaran, R. (1997). *Molecular architecture of polysaccharide helices in oriented fibers*. *Adv. Carbohydr. Chem. Biochem.* **52**, 311–439.

Chandrasekaran, R. & Arnott, S. (1989). *The structures of DNA and RNA helices in oriented fibers*. In *Landolt–Börnstein Numerical Data and Functional Relationships in Science and Technology*, Vol. VII/1b, edited by W. Saenger, pp. 31–170. Berlin: Springer-Verlag.

Chandrasekaran, R., Bian, W. & Okuyama, K. (1998). *Three-dimensional structure of guaran*. *Carbohydr. Res.* **312**, 219–224.

Chandrasekaran, R., Giacometti, A. & Arnott, S. (2000*a*). *Structure of poly d(T)·poly d(A)·poly d(T)*. *J. Biomol. Struct. Dynam.* **17**, 1011–1022.

Chandrasekaran, R., Giacometti, A. & Arnott, S. (2000*b*). *Structure of poly (U)·poly (A)·poly (U)*. *J. Biomol. Struct. Dynam.* **17**, 1023–1034.

Chandrasekaran, R., Giacometti, A. & Arnott, S. (2000*c*). *Structure of poly (I)·poly (A)·poly (I)*. *J. Biomol. Struct. Dynam.* **17**, 1035–1045.

Chandrasekaran, R., Puigjaner, L. C., Joyce, K. L. & Arnott, S. (1988). *Cation interactions in gellan: an X-ray study of the potassium salt*. *Carbohydr. Res.* **181**, 23–40.

Chandrasekaran, R. & Radha, A. (1992). *Structure of poly d(A)·poly d(T)*. *J. Biomol. Struct. Dynam.* **10**, 153–168.

Chandrasekaran, R., Radha, A. & Lee, E. J. (1994). *Structural roles of calcium ions and side chains in welan: an X-ray study*. *Carbohydr. Res.* **252**, 183–207.

Chandrasekaran, R., Radha, A., Lee, E. J. & Zhang, M. (1994). *Molecular architecture of araban, galactoglucan and welan*. *Carbohydr. Polym.* **25**, 235–243.

Chandrasekaran, R., Radha, A. & Park, H.-S. (1995). *Sodium ions and water molecules in the structure of poly d(A)·poly d(T)*. *Acta Cryst.* D**51**, 1025–1035.

Chandrasekaran, R., Radha, A. & Park, H.-S. (1997). *Structure of poly d(AI)·poly d(CT) in two different packing arrangements*. *J. Biomol. Struct. Dynam.* **15**, 285–305.

Cochran, W., Crick, F. H. & Vand, V. (1952). *The structure of synthetic polypeptides. I. The transform of atoms in a helix*. *Acta Cryst.* **5**, 581–586.

Cohen, C., Harrison, S. C. & Stephens, R. E. (1971). *X-ray diffraction from microtubules*. *J. Mol. Biol.* **59**, 375–380.

Franklin, R. E. & Gosling, R. G. (1953*a*). *Molecular configuration in sodium thymonucleate*. *Nature (London)*, **171**, 740–741.

Franklin, R. E. & Gosling, R. G. (1953*b*). *Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate*. *Nature (London)*, **172**, 156–157.

Franklin, R. E. & Klug, A. (1955). *The splitting of layer lines in X-ray fibre diagrams of helical structures: application to tobacco mosaic virus*. *Acta Cryst.* **8**, 777–780.

Fraser, R. D. B., MacRae, T. P., Miller, A. & Rowlands, R. J. (1976). *Digital processing of fibre diffraction patterns*. *J. Appl. Cryst.* **9**, 81–94.

Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). *An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules*. *J. Appl. Cryst.* **11**, 693–694.

Gonzalez, A., Nave, C. & Marvin, D. A. (1995). *Pf1 filamentous bacteriophage: refinement of a molecular model by simulated annealing using 3.3 Å resolution X-ray fibre diffraction data*. *Acta Cryst.* D**51**, 792–804.

Gregory, J. & Holmes, K. C. (1965). *Methods of preparing orientated tobacco mosaic virus sols for X-ray diffraction*. *J. Mol. Biol.* **13**, 796–801.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor*. *Acta Cryst.* **18**, 502–510.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures*. *Methods Enzymol.* **115**, 252–270.

Holmes, K. C., Popp, D., Gebhard, W. & Kabsch, W. (1990). *Atomic model of the actin filament*. *Nature (London)*, **347**, 44–49.

Inouye, H., Bond, J., Baldwin, M. A., Ball, H. L., Prusiner, S. B. & Kirschner, D. A. (2000). *Structural changes in a hydrophobic domain of the prion protein induced by hydration and by Ala→Val and Pro→Leu substitutions*. *J. Mol. Biol.* **300**, 1283–1296.

Inouye, H., Fraser, P. E. & Kirschner, D. A. (1993). *Structure of β-crystallite assemblies formed by Alzheimer β-amyloid protein analogs – analysis by X-ray diffraction*. *Biophys. J.* **64**, 502–519.

Ivanova, M. I. & Makowski, L. (1998). *Iterative low-pass filtering for estimation of the background in fiber diffraction patterns*. *Acta Cryst.* A**54**, 626–631.

Janaswamy, S. & Chandrasekaran, R. (2001). *Three-dimensional structure of the sodium salt of iota-carrageenan*. *Carbohydr. Res.* **335**, 181–194.

Janaswamy, S. & Chandrasekaran, R. (2002). *Effect of calcium ions on the organization of iota-carrageenan helices: an X-ray investigation*. *Carbohydr. Res.* **337**, 523–535.

Janaswamy, S. & Chandrasekaran, R. (2005). *Cation induced polymorphism in iota-carrageenan*. *Carbohydr. Polym.* **60**, 499–505.

Janaswamy, S. & Chandrasekaran, R. (2006). *Sodium ι-carrageenan: A paradigm of polymorphism and pseudopolymorphism*. *Macromolecules*, **39**, 3345–3349.

Janaswamy, S. & Chandrasekaran, R. (2008). *Heterogeneity in iota-carrageenan molecular structure: insights for polymorph II→III transition in the presence of calcium ions*. *Carbohydr. Res.* **343**, 364–373.

Kahn, R., Fourme, R., Gadet, A., Janin, J., Dumas, C. & André, D. (1982). *Macromolecular crystallography with synchrotron radiation: photographic data collection and polarization correction*. *J. Appl. Cryst.* **15**, 330–337.

Kendall, A., McDonald, M., Bian, W., Bowles, T., Baumgarten, S. C., Shi, J., Stewart, P. L., Bullitt, E., Gore, D., Irving, T. C., Havens, W. M., Ghabrial, S. A., Wall, J. S. & Stubbs, G. (2008). *Structure of flexible filamentous plant viruses*. *J. Virol.* **82**, 9546–9554.

Kendall, A., McDonald, M. & Stubbs, G. (2007). *Precise determination of the helical repeat of tobacco mosaic virus*. *Virology*, **369**, 226–227.

Kendall, A. & Stubbs, G. (2006). *Oriented sols for fiber diffraction from limited quantities or hazardous materials*. *J. Appl. Cryst.* **39**, 39–41

Kishimoto, A., Hasegawa, K., Suzuki, H., Taguchi, H., Namba, K. & Yoshida, M. (2004). *Beta-helix is a likely core structure of yeast prion Sup35 amyloid fibers*. *Biochem. Biophys. Res. Commun.* **315**, 739–745.

Klug, A., Crick, F. H. C. & Wyckoff, H. W. (1958). *Diffraction from helical structures*. *Acta Cryst.* **11**, 199–212.

Lorenz, M. & Holmes, K. C. (1993). *Computer processing and analysis of X-ray fiber diffraction data*. *J. Appl. Cryst.* **26**, 82–91.

McDonald, M., Kendall, A., Tanaka, M., Weissman, J. S. & Stubbs, G. (2008). *Enclosed chambers for humidity control and sample containment in fiber diffraction*. *J. Appl. Cryst.* **41**, 206–209.

MacGillavry, C. H. & Bruins, E. M. (1948). *On the Patterson transforms of fibre diagrams*. *Acta Cryst.* **1**, 156–158.

Makin, O. S. & Serpell, L. C. (2005). *Structures for amyloid fibrils*. *FEBS J.* **272**, 5950–5961.

Makowski, L. (1978). *Processing of X-ray diffraction data from partially oriented specimens*. *J. Appl. Cryst.* **11**, 273–283.

Makowski, L., Caspar, D. L. D. & Marvin, D. A. (1980). *Filamentous bacteriophage Pf1 structure determined at 7 Å resolution by refinement of models for the α-helical subunit*. *J. Mol. Biol.* **140**, 149–181.

Malinchik, S. B., Inouye, H., Szumowski, K. E. & Kirschner, D. A. (1998). *Structural analysis of Alzheimer's β(1–40) amyloid: protofilament assembly of tubular fibrils*. *Biophys. J.* **74**, 537–545.

Marvin, D. A., Wiseman, R. L. & Wachtel, E. J. (1974). *Filamentous bacterial viruses. XI. Molecular architecture of the class II (Pf1, Xf) virion*. *J. Mol. Biol.* **82**, 121–138.

Meyer, K. H. & Misch, L. (1937). *Positions des atomes dans le nouveau modèle spatial de la cellulose*. *Helv. Chim. Acta*, **20**, 232–244.

Millane, R. P. (1989). *R factors in X-ray fiber diffraction. II. Largest likely R factors*. *Acta Cryst.* A**45**, 573–576.

Millane, R. P. & Arnott, S. (1985). *Background removal in X-ray fiber diffraction patterns*. *J. Appl. Cryst.* **18**, 419–423.

Millane, R. P. & Arnott, S. (1986). *Digital processing of X-ray diffraction patterns from oriented fibers*. *J. Macromol. Sci. Phys. B*, **24**, 193–227.

Namba, K., Pattanayak, R. & Stubbs, G. (1989). *Visualization of protein–nucleic acid interactions in a virus: refinement of intact tobacco mosaic virus at 2.9 Å resolution by fiber diffraction data*. *J. Mol. Biol.* **208**, 307–325.

Namba, K. & Stubbs, G. (1985). *Solving the phase problem in fiber diffraction. Application to tobacco mosaic virus at 3.6 Å resolution*. *Acta Cryst.* A**41**, 252–262.

Namba, K. & Stubbs, G. (1987*a*). *Isomorphous replacement in fiber diffraction using limited numbers of heavy-atom derivatives*. *Acta Cryst.* A**43**, 64–69.

Namba, K. & Stubbs, G. (1987*b*). *Difference Fourier syntheses in fiber diffraction*. *Acta Cryst.* A**43**, 533–539.

Nambudripad, R., Stark, W. & Makowski, L. (1991). *Neutron diffraction studies of the structure of filamentous bacteriophage Pf1 – demonstration that the coat protein consists of a pair of α-helices with an intervening, non-helical loop*. *J. Mol. Biol.* **220**, 359–379.

Oda, T., Iwasa, M., Aihara, T., Maéda, Y. & Narita, A. (2009). *The nature of the globular- to fibrous-actin transition*. *Nature (London)*, **457**, 441–445.

Pauling, L. & Corey, R. B. (1951). *Atomic coordinates and structure factors for two helical configurations of polypeptide chains*. *Proc. Natl Acad. Sci. USA*, **37**, 235–240.

Pauling, L. & Corey, R. B. (1953). *Two rippled-sheet configurations of polypeptide chains, and a note about the pleated sheets*. *Proc. Natl Acad. Sci. USA*, **39**, 253–256.

Rajkumar, G., AL-Khayat, H., Eakins, F., He, A., Knupp, C. & Squire, J. (2005). *FibreFix – a new integrated CCP13 software package*. *Fibre Diffr. Rev.* **13**, 11–18.

Ramachandran, G. N. & Kartha, G. (1955). *Structure of collagen*. *Nature (London)*, **176**, 593–595.

Rich, A. & Crick, F. H. C. (1955). *The structure of collagen*. *Nature (London)*, **176**, 915–916.

Shotton, M. W., Pope, L. H., Forsyth, V. T., Denny, R. C., Archer, J., Langan, P., Ye, H. & Boote, C. (1998). *New developments in instrumentation for X-ray and neutron fibre diffraction experiments*. *J. Appl. Cryst.* **31**, 758–766.

Shotton, M. W., Pope, L. H., Forsyth, T., Langan, P., Denny, R. C., Giesen, U., Dauvergne, M. T. & Fuller, W. (1997). *A high-angle neutron fibre diffraction study of the hydration of deuterated A-DNA*. *Biophys. Chem.* **69**, 85–96.

Smith, P. J. C. & Arnott, S. (1978). *LALS: a linked-atom least-squares reciprocal-space refinement system incorporating stereochemical restraints to supplement sparse diffraction data*. *Acta Cryst.* A**34**, 3–11.

Squire, J. M., Al-Khayat, H. A. & Yagi, N. (1993). *Muscle thin filament structure and regulation. Actin sub-domain movements and the tropomyosin shift modelled from low-angle X-ray diffraction*. *J. Chem. Soc. Faraday Trans.* **89**, 2717–2726.

Stroud, W. J. & Millane, R. P. (1995). *Analysis of disorder in biopolymer fibers*. *Acta Cryst.* A**51**, 790–800.

Stubbs, G. (1987). *The Patterson function in fiber diffraction*. In *Patterson and Pattersons*, edited by J. P. Glusker, E. K. Patterson & M. Rossi, pp. 548–557. New York: Oxford University Press.

Stubbs, G. (1989). *The probability distributions of X-ray intensities in fiber diffraction: largest likely values for fiber diffraction R factors*. *Acta Cryst.* A**45**, 254–258.

Stubbs, G. J. & Diamond, R. (1975). *The phase problem for cylindrically averaged diffraction patterns. Solution by isomorphous replacement and application to tobacco mosaic virus*. *Acta Cryst.* A**31**, 709–718.

Stubbs, G. & Makowski, L. (1982). *Coordinated use of isomorphous replacement and layer-line splitting in the phasing of fiber diffraction data*. *Acta Cryst.* A**38**, 417–425.

Stubbs, G., Namba, K. & Makowski, L. (1986). *Application of restrained least-squares refinement to fiber diffraction from macromolecular assemblies*. *Biophys. J.* **49**, 58–60.

Torbet, J. (1987). *Using magnetic orientation to study structure and assembly*. *Trends Biochem. Sci.* **12**, 327–330.

Tsuruta, H. & Irving, T. C. (2008). *Experimental approaches for solution X-ray scattering and fiber diffraction*. *Curr. Opin. Struct. Biol.* **18**, 601–608.

Walkinshaw, M. D. & Arnott, S. (1981). *Conformations and interactions of pectins. I. X-ray diffraction analysis of sodium pectate in neutral and acidified forms*. *J. Mol. Biol.* **153**, 1055–1073.

Wang, H., Culver, J. N. & Stubbs, G. (1997). *Structure of ribgrass mosaic virus at 2.9 Å resolution: evolution and taxonomy of tobamoviruses*. *J. Mol. Biol.* **269**, 769–779.

Wang, H. & Stubbs, G. (1993). *Molecular dynamics in refinement against fiber diffraction data*. *Acta Cryst.* A**49**, 504–513.

Wang, H. & Stubbs, G. (1994). *Structure determination of cucumber green mottle mosaic virus by X-ray fiber diffraction. Significance for the evolution of tobamoviruses*. *J. Mol. Biol.* **239**, 371–384.

Watson, J. D. & Crick, F. H. C. (1953). *Molecular structure of nucleic acids*. *Nature (London)*, **171**, 737–738.

Welsh, L. C., Symmons, M. F. & Marvin, D. A. (2000). *The molecular structure and structural transition of the α-helical capsid in filamentous bacteriophage Pf1*. *Acta Cryst.* D**56**, 137–150.

Welsh, L. C., Symmons, M. F., Sturtevant, J. M., Marvin, D. A. & Perham, R. N. (1998). *Structure of the capsid of Pf3 filamentous phage determined from X-ray fibre diffraction data at 3.1 Å resolution*. *J. Mol. Biol.* **283**, 155–177.

Winter, W. T., Smith, P. J. C. & Arnott, S. (1975). *Hyaluronic acid: structure of a fully extended 3-fold helical sodium salt and comparison with the less extended 4-fold helical forms*. *J. Mol. Biol.* **99**, 219–235.

Yamashita, I., Hasegawa, K., Suzuki, H., Vonderviszt, F., Mimori-Kiyosue, Y. & Namba, K. (1998). *Structure and switching of bacterial flagellar filaments studied by X-ray fiber diffraction*. *Nat. Struct. Biol.* **5**, 125–132.

Yamashita, I., Suzuki, H. & Namba, K. (1998). *Multiple-step method for making exceptionally well-oriented liquid-crystalline sols of macromolecular assemblies*. *J. Mol. Biol*. **278**, 609–615.

Yamashita, I., Vonderviszt, F., Mimori, Y., Suzuki, H., Oosawa, K. & Namba, K. (1995). *Radial mass analysis of the flagellar filament of Salmonella: implications for the subunit folding*. *J. Mol. Biol.* **253**, 547–558.

Zugenmaier, P. & Sarko, A. (1980). *The variable virtual bond modeling technique for solving polymer crystal structures*. *Am. Chem. Soc. Symp. Ser.* **141**, 225–237.

International Tables for Crystallography (2012). Vol. F, Chapter 19.6, pp. 593–614.

Chapter 19.6. Electron cryomicroscopy of biological macromolecules

T. S. Baker and R. Henderson

19.6.1. Abbreviations used

0D	Zero-dimensional (single particles)
1D	One-dimensional (helical)
2D	Two-dimensional
3D	Three-dimensional
EM	Electron microscope/microscopy
CryoTEM	Transmission electron cryomicroscopy
FEG	Field-emission gun
CTF	Contrast-transfer function
MTF	Modulation transfer function
DQE	Detective quantum efficiency
DPR	Differential phase residual
FSC	Fourier shell correlation
CCD	Charge-coupled device (slow-scan TV detector)

19.6.2. Introduction: macromolecular structure determination using electron microscopy

The two principal methods of macromolecular structure determination that use scattering techniques are electron microscopy and X-ray crystallography. The most important difference between the two is that the scattering cross section is about 10^5 times greater for electrons than it is for X-rays, so significant scattering using electrons is obtained for specimens that are 1 to 10 nm thick, whereas scattering or absorption of a similar fraction of an illuminating X-ray beam requires crystals that are 100 to 500 µm thick, though in recent years smaller crystals, still > 1 µm thick, have become usable at synchrotron sources. The second main difference is that electrons are much more easily focused than X-rays since they are charged particles that can be deflected by electric or magnetic fields. As a result, electron lenses are much superior to X-ray lenses and can be used to produce a magnified image of an object as easily as a diffraction pattern. This allows the electron microscope to be switched back and forth instantly between imaging and diffraction modes so that the image of a single molecule at any magnification can be obtained as conveniently as the electron-diffraction pattern of a thin crystal.

In the early years of electron microscopy of macromolecules, electron micrographs of molecules embedded in a thin film of heavy-atom stains (Brenner & Horne, 1959; Huxley & Zubay, 1960) were used to produce pictures which were interpreted directly. Beginning with the work of Klug (Klug & Berger, 1964), a more rigorous approach to image analysis led first to the interpretation of the two-dimensional (2D) images as the projected density summed along the direction of view and then to the ability to reconstruct the three-dimensional (3D) object from which the images arose (DeRosier & Klug, 1968; Hoppe *et al.*, 1968), with subsequent more sophisticated treatment of image contrast transfer (Erickson & Klug, 1971).

Later, macromolecules were examined by electron diffraction and imaging without the use of heavy-atom stains by embedding the specimens in either a thin film of glucose (Unwin & Henderson, 1975) or in a thin film of rapidly frozen water (Taylor & Glaeser, 1974; Dubochet *et al.*, 1982, 1988), which required the specimen to be cooled while it was examined in the electron microscope. This use of unstained specimens thus led to the structure determination of the molecules themselves, rather than the structure of a ‘negative stain’ excluding volume, and has created the burgeoning field of 3D electron microscopy of macromolecules.

Many medium-resolution (*i.e.*, 5–15 Å) structures of macromolecular assemblies (*e.g.* ribosomes), spherical and helical viruses, and larger protein molecules have now been determined by transmission electron cryomicroscopy (cryoTEM) in ice. A small number of atomic resolution structures have been obtained by cryoTEM of thin 2D crystals embedded in glucose, trehalose or tannic acid (Henderson *et al.*, 1990; Kühlbrandt *et al.*, 1994; Nogales *et al.*, 1998; Murata *et al.*, 2000; Gonen *et al.*, 2005; Holm *et al.*, 2006), where specimen cooling reduced the effect of radiation damage. One of these, the structure of bacteriorhodopsin (Henderson *et al.*, 1990) provided the first structure of a seven-helix membrane protein, and recent studies, particularly on aquaporins, are providing detailed insights into lipid–protein interactions (Reichow & Gonen, 2009). During the last few years, two different helical structures have allowed atomic models to be constructed from images of ice-embedded specimens (Miyazawa *et al.*, 2003; Yonekura, Maki-Yonekura & Namba, 2003) and the resolution of the best icosahedral structures has likewise improved dramatically to yield atomic models (*e.g.* Jiang *et al.*, 2008; Yu, Jin & Zhou, 2008; Zhang *et al.*, 2008; Chen *et al.*, 2009). Medium-resolution density distributions can often be interpreted in terms of the chemistry of the structure if a high-resolution model of one or more of the component pieces has already been obtained by X-ray, electron microscopy or NMR methods. As a result, the use of electron microscopy is becoming a powerful technique for which, in some cases, no alternative approach is possible. Numerous useful reviews (*e.g.* Amos *et al.*, 1982; Dubochet *et al.*, 1988; Baker *et al.*, 1999; van Heel *et al.*, 2000; Frank, 2002; Henderson, 2004; Chiu *et al.*, 2005) and books (Frank, 2006*b*; Glaeser *et al.*, 2007) have been written.

19.6.3. Physics of electron scattering and radiation damage

A schematic overview of scattering and imaging in the electron microscope is depicted in Fig. 19.6.3.1. The incident electron beam passes through the specimen and individual electrons are either unscattered or scattered by the atoms of the specimen. This scattering occurs either elastically, with no loss of energy and therefore no energy deposition in the specimen, or inelastically, with consequent energy loss by the scattered electron and accompanying energy deposition in the specimen, resulting in radiation damage. The electrons emerging from the specimen are then collected by the imaging optics, shown here for simplicity as a single lens, but in practice consisting of a complex system of five or six lenses with intermediate images being produced at successively higher magnification at different positions down the column. Finally, in the viewing area, either the electron-

Copyright © 2012 International Union of Crystallography

diffraction pattern or the image can be seen directly by eye on the phosphor screen, or detected by a TV or CCD camera, or recorded on photographic film.

In practice, the electron flux through the specimen is relatively low, so that only one electron at a time is in the column interacting with the specimen. This means that the unscattered and elastically scattered beams are in reality quantum-mechanical wavefronts, in which a wave function describes the probability distribution until the electron is actually recorded in the detector or on film. Where an inelastic event results in energy deposition and accompanying radiation damage in the specimen, the path of that electron, as well as its energy, wavelength and probability distribution, are slightly altered and this degrades the image.

19.6.3.1. Elastic and inelastic scattering

The coherent, elastically scattered electrons contain the high-resolution information describing the structure of the specimen. The amplitudes and phases of the scattered electron beams are directly related to the amplitudes and phases of the Fourier components of the atomic distribution in the specimen. When the scattered beams are recombined with the unscattered beam in the image, they create an interference pattern (the image) which, for thin specimens, is related approximately linearly to the density variations in the specimen. Information about the structure of the specimen can then be retrieved by digitization and computer-based image processing, as described in Section 19.6.5. The elastic scattering cross sections for electrons are not as simply related to the atomic composition as happens with X-rays. With X-ray diffraction, the scattering factors are simply proportional to the number of electrons in each atom, normally equal to the atomic number. Since elastically scattered electrons are in effect diffracted by the electrical potential inside atoms, the scattering factor for electrons depends not only on the nuclear charge but also on the size of the surrounding electron cloud, which screens the nuclear charge. As a result, electron scattering factors in the resolution range of interest in macromolecular structure determination (up to $\frac{1}{3}$ Å^{-1}) are very sensitive to the effective radius of the outer valency electrons and therefore depend sensitively on the chemistry of bonding. Although this is a fascinating field in itself with interesting work already carried out by the gas-phase electron-diffraction community (*e.g.* Hargittai & Hargittai, 1988), it is still an area where much work remains to be done. At present, it is probably adequate to think of the density obtained in macromolecular structure analysis by electron microscopy as roughly equivalent to the electron density obtained by X-ray diffraction but with the contribution from hydrogen atoms being somewhat greater relative to carbon, nitrogen and oxygen.

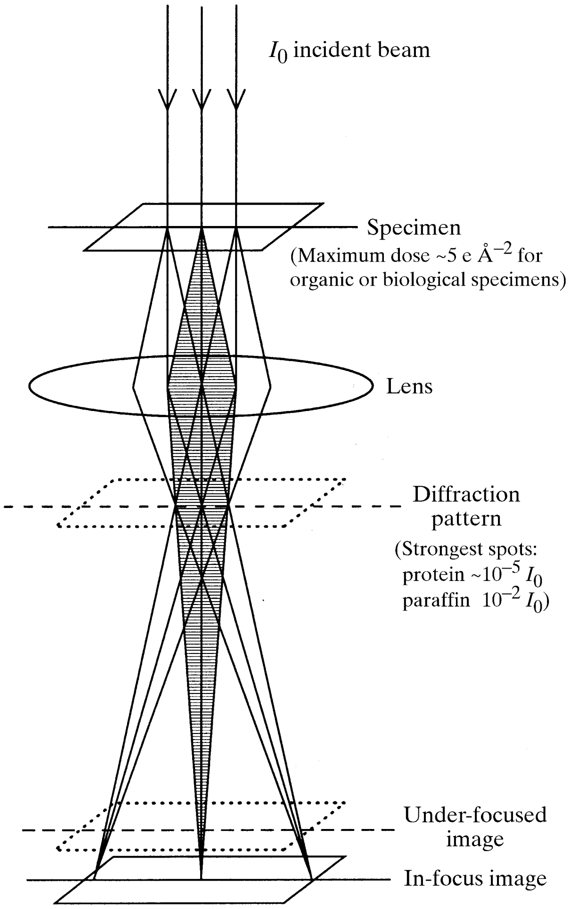

I_0 incident beam
Specimen
(Maximum dose ~5 e Å^{-2} for organic or biological specimens)
Lens
Diffraction pattern
(Strongest spots: protein ~10^{-5} I_0 paraffin 10^{-2} I_0)
Under-focused image
In-focus image

Figure 19.6.3.1
Schematic diagram showing the principle of image formation and diffraction in the transmission electron microscope. The incident beam, I_0, illuminates the specimen. Scattered and unscattered electrons are collected by the objective lens and focused back to form first an electron-diffraction pattern and then an image. For a 2D or 3D crystal, the electron-diffraction pattern would show a lattice of spots, each of whose intensity is a small fraction of that of the incident beam. In practice, an in-focus image has virtually no contrast, so images are recorded with the objective lens underfocused to take advantage of the out-of-focus phase-contrast mechanism. Higher electron doses than indicated can be used at lower resolution. The given dose of 5 e Å^{-2} applies for high-resolution (~4 Å) structural features.

Those electrons that are inelastically scattered lose energy to the specimen by a number of mechanisms. The energy-loss spectrum for a typical biological specimen is dominated by the large cross section for plasmon scattering in the energy range 20–30 eV with a continuum in the distribution which decreases up to higher energies. At discrete high energies, specific inner electrons in the *K* shell of carbon, nitrogen or oxygen can be ejected with corresponding peaks in the energy-loss spectrum appearing at 200–400 eV. Any of these inelastic interactions produces a deviation in the direction of the scattered electron and, as a result, the resolution of any information present in the energy-loss electron signal is reduced. Chromatic aberration in the objective lens results in these inelastic electrons contributing to the image only at relatively low resolution, around 15 Å (Isaacson *et al.*, 1974). Consequently, the inelastically scattered electrons are generally considered to contribute little, except noise, to the images. In the future, once electron cryo-microscopes with chromatic aberration correctors become available (expected by 2012), one may be able to obtain improved energy-loss images. Chromatic-aberration-corrected electron microscopes are already available for materials science (Kabius *et al.*, 2009).

19.6.3.2. Radiation damage

The most important consequence of inelastic scattering is the deposition of energy into the specimen. This is initially transferred to secondary electrons, which have an average energy (20 eV) that is five or ten times greater than the valence bond energies. These secondary electrons interact with other components of the specimen and produce numerous reactive chemical species, including free radicals. In ice-embedded samples, these would be predominantly highly reactive hydroxyl free radicals that arise from the frozen water molecules. In turn, these react with the embedded macromolecules and create a great variety of radiation products such as modified side chains, cleaved polypeptide backbones and a host of molecular fragments. From radiation-chemistry studies, it is known that thiol or disulfide groups react more quickly than aliphatic groups and that aromatic groups, including nucleic acid bases, are the most resistant (Burmeister, 2000; Weik *et al.*, 2000; Meents *et al.*, 2010). Nevertheless, the end effect of the inelastic scattering is the degradation of the specimen to produce a cascade of heterogeneous products, some of which resemble the starting structure more closely than others. Some of the secondary electrons also escape from the surface of the specimen, causing it to charge up during the exposure. As a rough rule, for 100 kV electrons the dose that can be used to produce an image in which the starting structure at high resolution is still recognizable is about 1 e Å^{-2} for organic or biological materials at room temperature, 5 e Å^{-2} for a specimen near liquid-nitrogen temperature ($-170\ ^\circ\text{C}$) and 10 e Å^{-2} for a specimen near liquid-helium temperature (4–8 K). However, individual experimenters will often exceed these doses if they wish to enhance the low-resolution information in the images, which is less sensitive to radiation damage. The effects of radiation damage owing to electron irradiation are essentially identical to those from X-ray or neutron irradiation for biological macromolecules except for the amount of energy deposition per useful coherent elastically scattered event (Henderson, 1995). For electrons scattered by biological structures at all electron energies of interest, the number of inelastic events exceeds the number of elastic events by a factor of three to four, so that 60 to 80 eV of energy is deposited for each elastically scattered electron. This limits the amount of information in an image of a single biological macromolecule. Consequently, the 3D atomic structure cannot be determined from a single molecule but requires the averaging of the information from a few thousand molecules in theory, and even more in practice (Henderson, 1995; Rosenthal & Henderson, 2003; Glaeser, 2008*b*; Cheng & Walz, 2009). Crystals used for X-ray or neutron diffraction contain many orders of magnitude more molecules.

It is possible to collect both the elastically and the inelastically scattered electrons simultaneously with an energy analyser and, if a fine electron beam is scanned over the specimen, then a scanning transmission electron micrograph displaying different properties of the specimen can be obtained (*e.g.* Tichelaar *et al.*, 1980; Crewe, 1983; Muller *et al.*, 2008). Alternatively, conventional transmission electron microscopes to which an energy filter has been added can be used to select out a certain energy band of the electrons from the image (*e.g.* Yonekura *et al.*, 2006; Maki-Yonekura & Yonekura, 2008). Both these types of microscopy can contribute in other ways to the knowledge of structure, but in this article we concentrate on high-voltage phase-contrast electron microscopy of unstained macromolecules most often embedded in ice, because this is the method of widest impact in structural biology.

19.6.3.3. *Required properties of the illuminating electron beam*

The important properties of the image in terms of defocus, astigmatism, and the presence and effect of amplitude or phase contrast are discussed in Section 19.6.5. Microscopes with field emission guns (FEGs) produce the best-quality incident electron beams. This is because the electrons from a FEG are emitted from a very small volume at the tip, which is the apparent source size. Once these electrons have been collected by the condenser lens and used to produce the illuminating beam, that beam of electrons is then very parallel (divergence of $\sim 10^{-2}$ mrad) and therefore spatially coherent. Similarly, because the emitting tip of a FEG is not heated as much as a conventional thermionic tungsten source, the thermal-energy spread of the electrons is relatively small (0.5 to 1.0 eV) and the illuminating beam is closer to being monochromatic. Electron beams can also be produced by a normal heated tungsten source, which gives a less parallel beam with a larger energy spread, but is nevertheless adequate for cryoTEM if the highest resolution images are not required.

19.6.4. Three-dimensional electron cryomicroscopy of macromolecules

The determination of 3D structures by cryoTEM methods follows a common scheme for all macromolecules (Fig. 19.6.4.1). A more detailed discussion of the individual steps as applied to different classes of macromolecules appears in subsequent sections. Briefly, each specimen must be prepared in a relatively homogeneous aqueous form (1D or 2D crystals or a suspension of single particles in a limited number of states) at relatively high concentration, rapidly frozen (vitrified) as a thin film, transferred into the electron microscope and imaged using low-dose selection and focusing procedures. The resulting images are more frequently being recorded using an electronic detector. Alternatively, if recorded on film, they must then be digitized. Digitized images are then processed using computer programs that allow different views of the specimen to be combined into a 3D reconstruction that can be interpreted in terms of other available structural, biochemical and molecular data.

19.6.4.1. *Overview of conceptual steps*

Radiation damage by the illuminating electron beam generally allows only one good picture (micrograph) to be obtained from

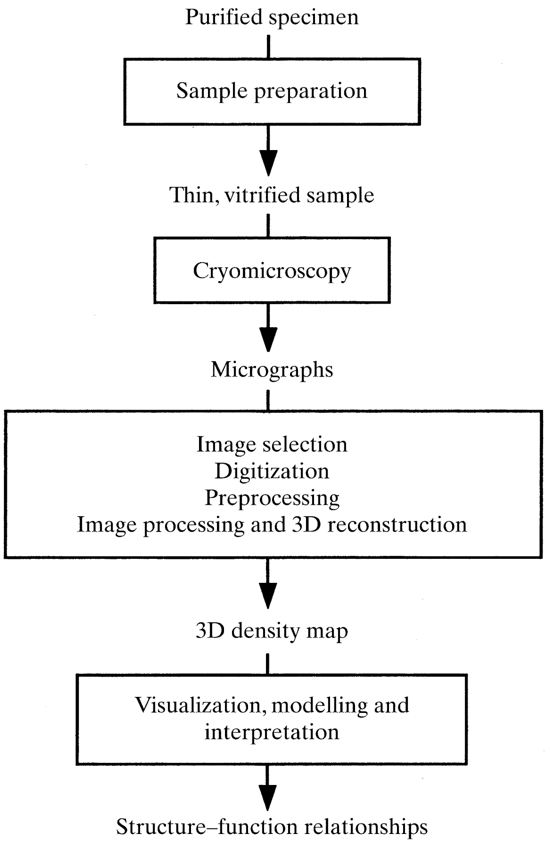


Figure 19.6.4.1
Flow diagram showing common procedures involved in cryoTEM from sample preparation to map interpretation.

each molecule or macromolecular assembly. In this micrograph, the signal-to-noise ratio of the 2D projection image is normally too small to determine the projected structure accurately. This implies firstly that it is necessary to average many images of different molecules taken from essentially the same viewpoint to increase the signal-to-noise ratio, and secondly that many of these averaged projections, taken from different directions, must be combined to build up the information necessary to determine the 3D structure of the molecule. Thus, the two key concepts are: (1) averaging to a greater or lesser extent depending on resolution, particle size and symmetry to increase the signal-to-noise ratio; and (2) the combination of different projections to build a 3D map of the structure.

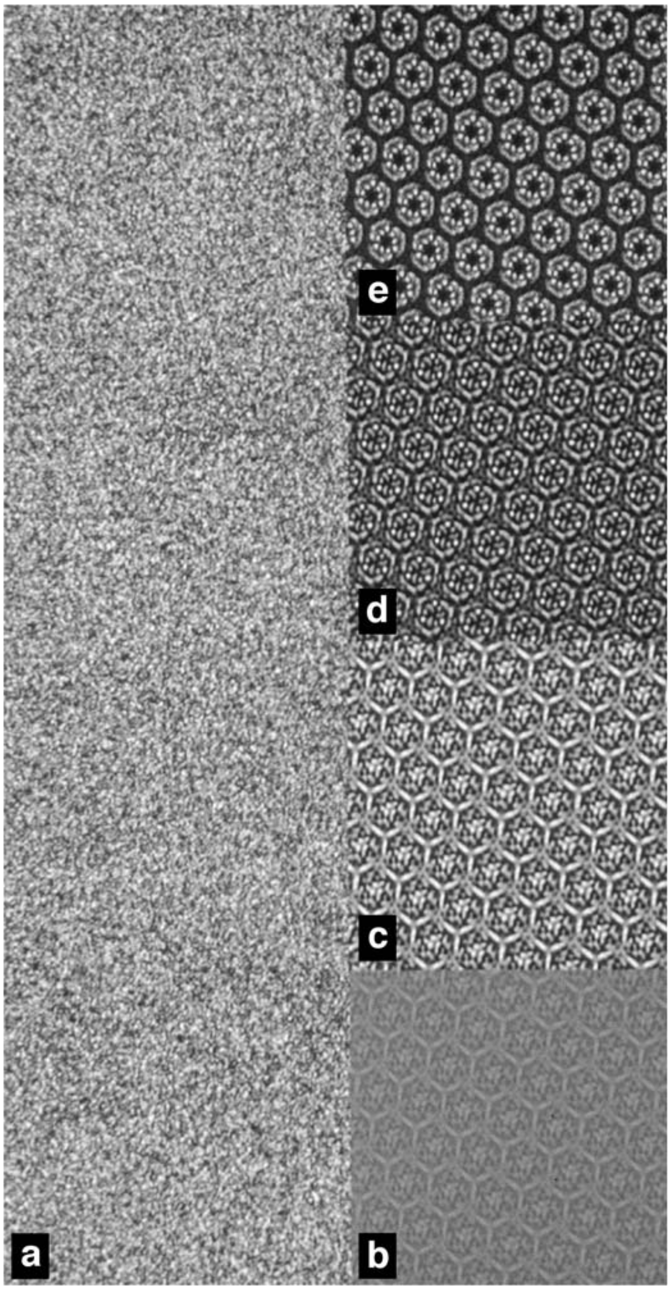


Figure 19.6.4.2
A display of the results at different stages of image processing of a digitized micrograph of a 2D crystal of bacteriorhodopsin. The left panel (*a*) shows an area of the raw digitized micrograph in which only electron noise is visible. The lower right panel (*b*) shows the results of the averaging of unit cells from the whole picture by unbending in real space and filtering in reciprocal space. The scale of the density in (*b*) is the same as that in the original micrograph, showing that the signal is very much weaker than the noise. Panel (*c*) shows the same density as in (*b*) but with contrast increased tenfold to show that the signal in the original picture is approximately 10× below the noise level. Panel (*d*) shows the density after correction for the contrast-transfer function (CTF) due in this case to a defocus of 6000 Å. Panel (*e*) shows the density after further threefold crystallographic averaging (the space group is *p*3 with a cell dimension of 62.5 Å) and replacement of image amplitudes by electron-diffraction amplitudes. Panel (*e*) therefore shows an almost perfect atomic resolution image of the projected structure of bacteriorhodopsin. The trimeric rings of molecules are centred on the crystallographic threefold axes and the internal structure shows α-helical segments in the protein.

In addition, there are various technical corrections that must be made to the image data to allow an unbiased model of the structure to be obtained. These include correction for the phase contrast-transfer function (CTF) of the microscope and, at high resolution, for the effects of beam tilt. For crystals, it is also possible to combine electron-diffraction amplitudes with image phases to produce a more accurate structure (Unwin & Henderson, 1975), or to use new electron-diffraction data to improve a previous model (Gonen *et al.*, 2005), and in general to correct for loss of high-resolution contrast for any reason by ‘sharpening’ the data by application of a negative temperature factor (Havelka *et al.*, 1995; Rosenthal & Henderson, 2003; Fernandez *et al.*, 2008).

The idea of increasing the signal-to-noise ratio in electron images of unstained biological macromolecules by averaging was discussed in 1971 (Glaeser, 1971) and demonstrated in 1975 (Henderson & Unwin, 1975; Unwin & Henderson, 1975), though earlier work on stained specimens had shown the value of averaging to increase the signal-to-noise ratio (Klug & DeRosier, 1966). The improvement obtained, as in all repeated measurements, gives a factor of $N^{1/2}$ improvement in signal-to-noise ratio, where N is the number of times the measurement is made. The effect of averaging to produce an improvement in signal-to-noise ratio is seen most clearly in the processing of images from 2D crystals. For example, Fig. 19.6.4.2 shows the results of applying a sequence of corrections, beginning with averaging, to 2D crystals of bacteriorhodopsin in 2D space group *p*3. The panels show: (*a*) raw, digitized image, (*b*, *c*) 2D averaging, (*d*) correction for the microscope CTF, and (*e*) threefold crystallographic symmetry averaging of the phases and combination with electron-diffraction amplitudes. At each stage, the projected picture of the molecules gets clearer. The final stage results in a virtually noise-free projected structure for the molecule at near-atomic (3 Å) resolution.

The earliest successful application of the idea of combining projections to reconstruct the 3D structure of a biological assembly was made by DeRosier & Klug (1968). The idea is that each 2D projection corresponds after Fourier transformation to a central section of the 3D transform of the assembly. If enough independent projections are obtained, then the 3D transform will have been fully sampled and the structure can then be obtained by back-transformation of the averaged, interpolated and smoothed 3D transform. This procedure is shown schematically for a 3D object in the shape of a duck, which represents the molecule whose structure is being determined (Fig. 19.6.4.3).

In practice, the implementation of these concepts has been carried out in a variety of ways, since the experimental strategy and type of computer analysis used depends on the type of specimen, especially the molecular mass of the individual molecule, its symmetry and whether or not it assembles into an aggregate with 1D, 2D or 3D periodic order.

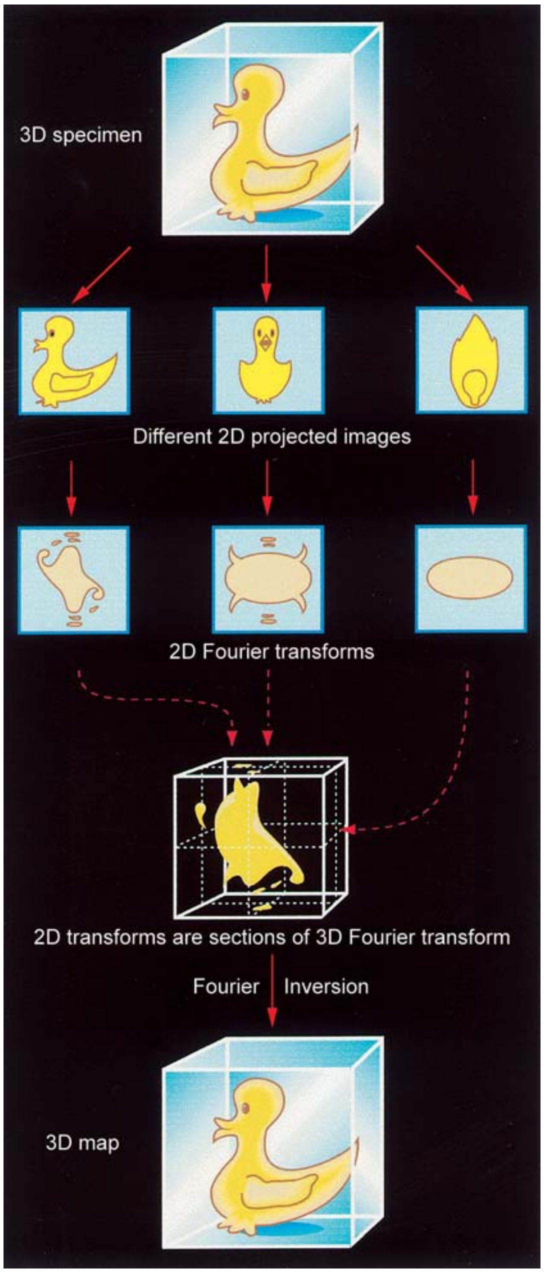

3D specimen
Different 2D projected images
2D Fourier transforms
2D transforms are sections of 3D Fourier transform
Fourier
Inversion
3D map

Figure 19.6.4.3
Schematic diagram to illustrate the principle of 3D reconstruction. Each 2D projected image, as recorded on the micrograph and after CTF correction, represents a section through the 3D Fourier transform. This is called the projection theorem. After accumulation of enough information from enough different views, a 3D map of the structure can be calculated by Fourier inversion.

19.6.4.2. Classification of macromolecules

The symmetry of a macromolecule or supramolecular complex is the primary determinant of how specimen preparation, microscopy and 3D image reconstruction are performed (Sections 19.6.4.3, 19.6.4.4 and 19.6.5). The classification of molecules according to their level of periodic order and symmetry (Table 19.6.4.1) provides a logical and convenient way to consider the means by which specimens are studied in 3D by microscopy.

Each type of specimen offers a unique set of challenges in obtaining 3D structural information at the highest possible resolution. The best resolutions achieved by 3D EM methods to date, at about 2–4 Å, have been obtained with several thin 2D crystals, in large part due to their excellent order, with helical arrays and icosahedral single particles now crossing the 4 Å barrier (Cheng & Walz, 2009).

With the exception of true 3D crystals, which must be sectioned to make them thin enough to study by transmission electron microscopy, the resolution obtained with biological specimens is generally dictated by the preservation of periodic order, and the symmetry and complexity of the object. Hence, studies for example of 2D crystals of the aquaporin membrane channel (Gonen *et al.*, 2005), helical acetylcholine receptor tubes (Miyazawa *et al.*, 2003) and bacterial flagella (Yonekura, Maki-Yonekura & Namba, 2003), three icosahedral virus capsids (Yu, Jin & Zhou, 2008; Zhang *et al.*, 2008; Chen *et al.*, 2009), the 70S *E. coli* ribosome (LeBarron *et al.*, 2008), and HIV trimers obtained by sub-tomogram averaging (Liu, Bartesaghi *et al.*, 2008), have yielded 3D density maps at resolutions of 1.9, 4.0, 4.0, 4.0, 3.8, 3.8, 6.7 and 19 Å, respectively.

If high resolution were the sole objective of EM, it would be necessary, given the capabilities of existing technology, to try to form well ordered 2D crystals, helical assemblies or icosahedral particles of each macromolecule of interest. Indeed, a number of different crystallization techniques have been devised (*e.g.* Yoshimura *et al.*, 1990; Kornberg & Darst, 1991; Jap *et al.*, 1992; Kubalek *et al.*, 1994; Hasler *et al.*, 1998; Wilson-Kubalek *et al.*, 1998; Dang *et al.*, 2005; Schmidt-Krey, 2007) and some of these have yielded new structural information about otherwise recalcitrant molecules like RNA polymerase (Polyakov *et al.*, 1998). However, despite the obvious technological advantages of having a molecule present in a highly ordered form, most macromolecules function not as highly ordered crystals or helices but instead as single particles (*e.g.* many enzymes) or, more likely, in concert with other macromolecules as occurs in supramolecular assemblies. Also, crystallization tends to constrain the number of conformational states a molecule can adopt and the crystal conformation might not be functionally relevant. Hence, though resolution may be restricted to less than that realized in the bulk of current X-ray crystallographic studies, cryoTEM methods provide a powerful means to study molecules that resist crystallization in 1D, 2D or 3D. These methods allow one to explore the dynamic events, different conformational states (as induced, for example, by altering the microenvironment of the specimen) and macromolecular interactions that are the key to understanding how each macromolecule functions.

19.6.4.3. Specimen preparation

The goal in preparing specimens for cryoTEM is to keep the biological sample in its native state as closely as possible in order to preserve the structure to atomic or near-atomic resolution in the microscope and during microscopy. The methods by which numerous types of macromolecules and macromolecular complexes have been prepared for cryoTEM studies are now well established (Adrian *et al.*, 1984; Bellare *et al.*, 1988; Dubochet *et al.*, 1988; Grassucci *et al.*, 2007). Most such methods involve cooling samples at a rate fast enough to permit vitrification (to a solid glass-like state) rather than crystallization of the bulk water. Noncrystalline biological macromolecules are typically vitrified by applying a small (often <5 µl) aliquot of a purified ~0.2–5 mg ml^{-1} suspension of sample to an EM grid coated with a carbon or holey carbon support film. The grid, secured with a pair of forceps and suspended over a container of ethane or propane cryogen slush (maintained near its freezing point by a reservoir of liquid nitrogen), is blotted nearly dry with a piece of filter paper. The grid is then plunged into the cryogen, and the sample, if thin

Table 19.6.4.1

Classification of macromolecules according to periodic order and symmetry

Periodic order	Type	Symmetry	Example macromolecule/complex	Representative reference
0D	Point group	C_1	70S ribosome (bacterial)	Villa *et al.* (2009)
		C_1	80S ribosome (mammalian)	Chandramouli *et al.* (2008)
		C_1	Centriole	Kenney *et al.* (1997)
		C_1	ATP synthase (mitochondrial)	Rubinstein *et al.* (2003)
		C_1	ATP synthase (yeast)	Lau *et al.* (2008)
		C_1	Yeast RNA polymerase III	Fernandez-Tornero *et al.* (2007)
		C_1	P-ATPase (*E. coli*)	Heitkamp *et al.* (2009)
		C_1	V-ATPase (bovine brain)	Gregorini *et al.* (2007)
		C_1	V-ATPase (*Manduca sexta*)	Muench *et al.* (2009)
		C_1	Bacteriophage φ29 virion/ghost	Tang *et al.* (2008)
		C_2	Transferrin–transferrin-receptor complex	Cheng *et al.* (2004)
		C_2	Gp130–receptor complex	Matadeen *et al.* (2007)
		C_2	Voltage-gated calcium channel	Wang *et al.* (2004)
		C_3	Mg-chelatase ID complex	Elmlund *et al.* (2008)
		C_4	Ryanodine receptor channel	Ludtke *et al.* (2005)
		C_4	CLIC2–ryanodine-receptor complex	Meng *et al.* (2009)
		C_5	Bacteriophage φ29 head	Tao *et al.* (1998)
		C_5	Hemocyanin (*Nautilus pompilius*)	Gatsogiannis *et al.* (2007)
		C_6	Plant plasma membrane H^+-ATPase	Ottmann *et al.* (2007)
		C_7	GroEL–GroES complexes	Ranson *et al.* (2006)
		C_7	Cytolysin (*Vibrio cholerae*)	He & Olson (2010)
		C_8	Outer membrane protein Wza (*E. coli*)	Ford *et al.* (2009)
		C_{12}	Bacteriophage T7 connector	Agirrezabala *et al.* (2005)
		C_{17}	TMV disc	Bloomer *et al.* (1978)
		C_{20}	Type III secretion injectisome	Marlovits *et al.* (2006)
		D_2	Phosphorylase kinase holoenzyme	Venien-Bryan *et al.* (2009)
		D_5	Hemocyanin (*Rapana thomasiana*)	Cheng *et al.* (2006)
		D_6	Clathrin coats	Fotin *et al.* (2004)
		D_6	Haemoglobin (*Lumbricus terrestris*)	Schatz *et al.* (1995)
		D_7	GroEL	Ludtke *et al.* (2008)
		D_7	20S proteosome	Rabl *et al.* (2008)
		D_8	Double aptosome	Yu *et al.* (2006)
		D_{48}	Recombinant ribonucleoprotein vault	Mikyas *et al.* (2004)
		T	Dps protein	Grant *et al.* (1998)
		O	Pyruvate dehydrogenase core	Mattevi *et al.* (1992)
		I	COPII coat assembly	Stagg *et al.* (2008)
		I	Human PDC E2 and tE2 cores	Yu, Hiromasa *et al.* (2008)
		I	Tricorn protease	Walz *et al.* (1997)
		I	Icosahedral viruses	Baker *et al.* (1999)
1D	Screw axis (helical)†		F-actin–fimbrin/plastin ABD2 complex	Galkin *et al.* (2008)
			Acto-myosin filament	Milligan (1996)
			Acetylcholine receptor tubes	Miyazawa *et al.* (2003)
			Bacterial flagella	Yonekura, Maki-Yonekura & Namba (2003)
			Bacterial ParM filament	Orlova *et al.* (2007)
			Bacterial type IV pilus	Craig *et al.* (2006)
			Bacteriophage fd	Wang *et al.* (2006)
			F-BAR membrane tubule	Frost *et al.* (2008)
			Microtubule	Li *et al.* (2002)
			P-type ATPases	Pomfret *et al.* (2007)
			Tobacco mosaic virus	Sachse *et al.* (2007)
2D	2D space group (2D crystal)	$p12_1$	Tubulin sheet	Nogales *et al.* (1998)
		$p22_12_1$	Prostaglandin E synthase	Jegerschold *et al.* (2008)
		$p22_12_1$; $p6$	Microsomal glutathione transferase 1	Holm *et al.* (2006)
		$p3$	Bacteriorhodopsin membrane	Henderson *et al.* (1990)
		$p321$	Light-harvesting complex II	Kühlbrandt *et al.* (1994)
		$p42_12$	Aquaporin AQPO membrane	Gonen *et al.* (2005)
		$p6$	Gap junction membrane	Fleishman *et al.* (2004)
3D	3D space group (3D crystal)	$P2_12_12_1$	Myosin S1 protein crystal	Winkelmann *et al.* (1991)
		$P6_5$ or $P6_4$	Insect flight muscle	Liu *et al.* (2006)

† The symmetry of a helical structure is defined by an n_m screw axis, which combines a rotation of $2\pi/n$ radians about an axis followed by a translation of m/n of the repeat distance. Because many helical structures are polymorphic, a different n_m symmetry is needed to specify each polymorph. This designation can be confusing: for example, for tobacco mosaic virus $n_m = 16.333_{0.333}$, because the helical translational repeat consists of 49 subunits in three turns of the basic helix.

enough (~0.2 µm or less), is vitrified in millisecond or shorter time periods (Mayer & Astl, 1992; Berriman & Unwin, 1994; White *et al.*, 1998; Kasas *et al.*, 2003).

The ability to freeze samples on a timescale of milliseconds affords cryoTEM one of its unique and perhaps still under-utilized advantages: capturing and visualizing dynamic structural events that occur over time periods of a few milliseconds or longer. Several devices that allow samples to be perturbed in a variety of ways as they are plunged into a cryogen have been described (*e.g.* Subramaniam *et al.*, 1993; Berriman & Unwin, 1994; Siegel *et al.*, 1994; Trachtenberg, 1998; White *et al.*, 1998, 2003; Shaikh *et al.*, 2009), and there are now commercially available freeze-plungers (*e.g.* Iancu *et al.*, 2006; Melanson, 2009). Examples of the use of such devices include spraying acetylcholine onto its receptor to cause the receptor channel to open (Unwin, 1995), lowering the pH of an enveloped virus sample to initiate early events of viral fusion (Fuller *et al.*, 1995), inducing a temperature jump with a flash-tube system to study phase transitions in liposomes (Siegel & Epand, 1997), or mixing myosin S1 fragments with F-actin to examine the geometry of the cross-bridge powerstroke in muscle (Walker *et al.*, 1999).

Crystalline (2D) samples can fortunately often be prepared for cryoTEM by means of simpler procedures, and vitrification of the bulk water is not always essential to achieve success (Cyrklaff & Kühlbrandt, 1994). Such specimens may be applied to the carbon film on an EM grid by normal adhesion methods, washed with 1–2% solutions of solutes like glucose, trehalose or tannic acid, blotted gently with filter paper to remove excess solution, air dried, loaded into a cold holder, inserted into the microscope, and, finally, cooled to liquid-nitrogen temperature.

19.6.4.4. Microscopy

Once the vitrified specimen is inserted into the microscope and sufficient time is allowed (~15 min) for the specimen stage to stabilize to minimize drift and vibration, microscopy is performed to generate a set of images that, with suitable processing procedures, can later be used to produce a reliable 3D reconstruction of the specimen at the highest possible resolution. To achieve this goal, imaging must be performed at an electron dose that minimizes beam-induced radiation damage to the specimen, with the objective lens of the microscope defocused to enhance phase contrast from the weakly scattering, unstained biological specimen, and under conditions that keep the specimen below the devitrification temperature (~−140 °C) and minimize its contamination.

The microscopist locates specimen areas suitable for image recording by searching the EM grid at very low magnification ($\leq 3000\times$) to keep the irradiation level quite low (<0.05 e Å^{-2} s^{-1}) while assessing sample quality. In microscopes operated at 200 keV or higher, where image contrast is very weak, it is helpful to perform the search procedure with the assistance of a CCD camera or a video-rate TV-intensified camera system. Indeed, these are now much more sensitive than looking at a phosphor screen by eye. For some specimens, like thin 2D crystals, searching is conveniently performed while viewing the low-magnification high-contrast image produced by using the diffraction lens to give a slightly defocused electron-diffraction pattern. Electronic detectors also have automatic histogram expansion settings that allow amplification of small contrast differences on a display monitor. These become more-or-less essential when using high-voltage (300 keV) microscopes.

After a desired specimen area is identified, the microscope is switched to high-magnification mode for focusing and astigmatism correction. These adjustments are typically performed in a region ~2–5 µm away from the chosen area at the same or higher magnification than that used for final image recording. The choice of magnification, defocus level, accelerating voltage, beam coherence, electron dose and other operating conditions is dictated by several factors. The most significant ones are the size of the particle or crystal unit cell being studied, the anticipated resolution of the images and the requirements of the image processing needed to compute a 3D reconstruction to the desired resolution. For most specimens at required resolutions from 3 to 30 Å, images are typically recorded at 25 000–50 000× magnification with an electron dose of between 5 and 20 e Å^{-2}. These conditions yield micrographs of sufficient optical density (film OD 0.2–1.5) or dynamic range (CCD detector) and image resolution for subsequent image-processing steps (Section 19.6.5). All modern EMs provide some mode of low-dose operation for imaging beam-sensitive, vitrified biological specimens.

The intrinsic low contrast of unstained specimens makes it impossible to observe and focus on specimen details directly as is routine with stained or metal-shadowed specimens. Focusing, aimed to enhance phase contrast in the recorded images while minimizing beam damage to the desired area, is achieved by judicious defocusing on a region that is adjacent to the region to be photographed and preferably situated on the microscope tilt axis. The appropriate focus level is set by adjusting the appearance of either the Fresnel fringes that occur at the edges of holes in the carbon film or the 'phase granularity' from the carbon support film.

Unfortunately, electron images do not give a direct rendering of the specimen density distribution. The relationship between image and specimen is described by the CTF which is characteristic of the particular microscope used, the specimen and the conditions of imaging. The microscope CTF arises from the objective-lens focal setting and from the spherical aberration present in all electromagnetic lenses, and varies with the defocus and accelerating voltage according to equation (19.6.4.1), an expression that includes both phase and amplitude contrast components. First, however, it is useful to consider briefly the essentials of amplitude contrast and phase contrast, two concepts carried over from optical microscopy. Amplitude contrast refers to the nature of the contrast in an image of an object that absorbs the incident illumination or scatters it in any other way so that a portion of it is lost. As a result, the image appears darker where greater absorption occurs. Phase contrast is required if an object is transparent (*i.e.* it is a pure phase object) and does not absorb but only scatters the incident illumination. Biological specimens for cryoTEM are almost pure phase objects and the scattering is relatively weak. Hence, the simple theory of image formation by a weak phase object applies (Reimer & Kohl, 2008; Spence, 2009). An exactly in-focus image of a phase object has no contrast variation since all the scattered illumination is focused back to equivalent points in the image of the object from which it was scattered. In optical microscopy, the use of a quarter wave plate can retard the phase of the direct unscattered beam, so that an in-focus image of a phase object has very high 'Zernicke' phase contrast (Glaeser, 2008*a*). However, there is as yet no simple quarter wave plate for electrons, so instead phase contrast is created by introducing phase shifts into the diffracted beams by adjustment of the excitation of the objective lens so that the image is slightly defocused. In addition, since all matter is composed of atoms and the electric potential inside each atom is very high near the nucleus, even the electron-scattering behaviour of the light atoms found in biological molecules deviates from that of a weak phase object. For a more in-depth discussion of this the reader should refer to Reimer & Kohl (2008) or Spence (2009). In practice, the proportion of 'amplitude' contrast is about 7% at 100 kV, 5% at 200 kV and 4% at 300 kV for low-dose images of protein molecules embedded in ice.

The overall dependence of the CTF on resolution, wavelength, defocus and spherical aberration is

$$\mathrm{CTF}(\nu) = -\Big\{\big(1 - F_{\rm amp}^2\big)^{1/2}\sin[\chi(\nu)] + F_{\rm amp}\cos[\chi(\nu)]\Big\}, \tag{19.6.4.1}$$

where $\chi(\nu) = \pi\lambda\nu^2(\Delta f - 0.5C_s\lambda^2\nu^2)$, ν is the spatial frequency (in Å^{-1}), $F_{\rm amp}$ is the fraction of amplitude contrast, λ is the electron wavelength (in Å), where $\lambda = 12.3/(V + 0.000000978V^2)^{1/2}$ (= 0.037, 0.025 and 0.020 Å for 100, 200 and 300 keV electrons, respectively), V is the voltage (in volts), Δf is the underfocus (in Å) and C_s is the spherical aberration of the objective lens of the microscope (in Å).

In addition, the CTF is attenuated by an envelope or damping function that depends upon the spatial and temporal coherence of the beam, specimen drift and other factors (Erickson & Klug, 1971; Wade & Frank, 1977; Wade, 1992). Representative CTFs for different amounts of defocus on a conventional tungsten and a FEG microscope are illustrated in Fig. 19.6.4.4. For a particular defocus setting of the objective lens, phase contrast in the electron image is positive and maximal only at a few specific spatial frequencies. Contrast is either lower than maximal, completely absent or it is opposite (inverted or reversed) from that at other frequencies. Hence, as the objective lens is focused, the electron microscopist selectively accentuates image details of a particular size.

Images are typically recorded 0.8–3.0 µm underfocus to enhance specimen features in the 20–40 Å size range and thereby facilitate phase-origin and specimen-orientation search procedures carried out in the image-processing steps (Section 19.6.5). However, this level of underfocus also enhances contrast in lower-resolution maps, which may help in interpretation. To obtain results at sub-nanometre (<10 Å) resolution, it is normal to record, process and combine data from several micrographs that span a range of defocus levels (*e.g.* Unwin & Henderson, 1975; Böttcher *et al.*, 1997). This strategy assures good information transfer at all spatial frequencies up to the limiting resolution but requires careful compensation for the effects of the microscope CTF during image processing. Also, the recording of image focal pairs or focal series from a given specimen area can be beneficial in determining origin and orientation parameters for processing of images of single particles (*e.g.* Cheng *et al.*, 1992; Conway & Steven, 1999).

Most high-resolution cryoTEM studies are now performed with microscopes operated at 200 keV or higher and with FEG electron sources (*e.g.* Zemlin, 1992; Zhou & Chiu, 1993; Ludtke *et al.*, 2008; Yu, Jin & Zhou, 2008; Zhang *et al.*, 2008). The high coherence of a FEG source ensures that phase contrast in the images remains strong out to high spatial frequencies (>1/3.5 Å^{-1}) even for highly defocused images. The use of higher voltages provides potentially higher resolution (greater depth of field – *i.e.* less curvature of the Ewald sphere – owing to the smaller electron-beam wavelength), better beam penetration (less multiple scattering), reduced problems with specimen charging that plague microscopy of unstained or uncoated vitrified specimens (Brink *et al.*, 1998) and reduced phase shifts associated with beam tilt.

Images are recorded on photographic film or on an electronic camera with either flood-beam or spot-scan procedures. Film, with its advantages of low cost, large field of view and high resolution (~10 µm), has remained the primary image-recording medium for most high-resolution cryoTEM applications, despite disadvantages of high background fog and need for chemical development and digitization. Electronic (usually phosphor/fibre optic/CCD) cameras provide image data directly in digital form and with very low background noise, but suffer from higher cost, limited field of view, limited spatial resolution caused by poor point-spread characteristics and a fixed pixel size (typically between 14 and 24 µm). They are useful, for example, for precise focusing and adjustment of astigmatism (*e.g.* Krivanek & Mooney, 1993; Sherman *et al.*, 1996) and are increasingly used to acquire image data for 3D reconstruction studies (*e.g.* Booth *et al.*, 2004; Sander *et al.*, 2005; Booth *et al.*, 2006; Chen *et al.*, 2008). There is a useful comparison of the MTF and DQE at different spatial frequencies between film and several electronic detectors (McMullan, Chen *et al.*, 2009). In summary, at present, all image recording is better at low (*e.g.* 120 kV) voltage with little to choose between different recording media, but film, in 2009, is still best at 300 keV.

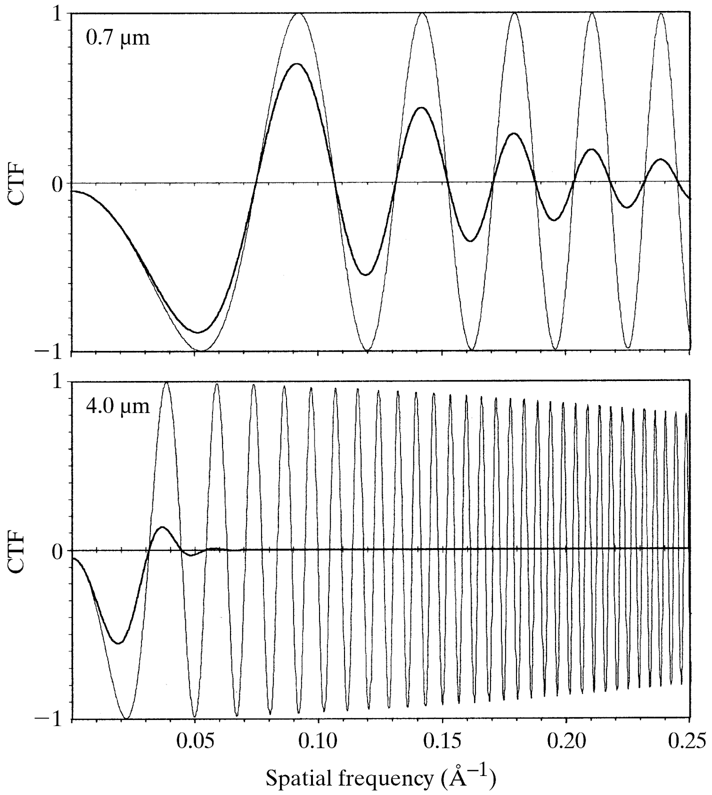


Figure 19.6.4.4
Representative plots of the microscope CTF as a function of spatial frequency, for two different defocus settings (0.7 and 4.0 µm underfocus) and for a field-emission (light curve) or tungsten (dark curve) electron source. All plots correspond to electron images formed in an electron microscope operated at 200 kV with objective-lens aberration coefficients $C_s = C_c = 2.0$ mm and assuming amplitude contrast of 4.8% (Toyoshima *et al.*, 1993). The spatial coherence, which is related to the electron source size and expressed as β, the half-angle of illumination, for tungsten and FEG electron sources was fixed at 0.3 and 0.015 mrad, respectively. Likewise, the temporal coherence (expressed as ΔE, the energy spread) was fixed at 1.6 and 0.5 eV for tungsten and FEG sources. The combined effects of the poorer spatial and temporal coherence of the tungsten source leads to a significant dampening, and hence loss of contrast, of the CTF at progressively higher resolutions compared with that observed in FEG-equipped microscopes. The greater number of contrast reversals with higher defocus arises because of the greater out-of-focus phase shifts.

For studies in which specimens must be tilted to collect 3D data, such as with 2D crystals, single particles that adopt preferred orientations on the EM grid or specimens requiring tomography, microscopy is performed in essentially the same way as described above. However, the limited tilt range (±60–70°) of most microscope goniometers can lead to non-isotropic resolution in the 3D reconstructions (the 'missing wedge' or missing cone' problem), and tilting generates a constantly varying defocus across the field of view in a direction normal to the tilt axis. The effects caused by this varying defocus level must be corrected in high-resolution applications (Henderson *et al.*, 1990).

19.6.4.5. Selection and preprocessing of digitized images

Before any image analysis or classification of the molecular images can be done, a certain amount of preliminary checking and normalization is required to ensure there is a reasonable chance that a homogeneous population of molecular images has been obtained. First, good-quality micrographs are selected in which the electron exposure is correct, there is negligible image drift or blurring, and there is minimal astigmatism and a reasonable amount of defocus to produce good phase contrast. This is usually done by visual examination and optical diffraction or examination of the computed diffraction pattern.

Once the best micrographs have been chosen, if they are on film they must be scanned and digitized on a suitable densitometer. The sizes of the steps between digitization of optical density, and the size of the sample aperture over which the optical density is averaged by the densitometer, must be sufficiently small to sample the detail present in the image at fine enough intervals (DeRosier & Moore, 1970). Normally, a circular (or square) sample aperture of diameter (or length of side) equal to the step between digitization is used. This avoids digitizing overlapping points, without missing any of the information recorded in the image. The size of the sample aperture and digitization step depends on the magnification selected and the resolution required. A value of one-quarter to one-third of the required limit of resolution (measured in μm on the emulsion) is normally ideal, since it avoids having too many data points (and therefore wasting computer resources) without losing anything during the measurement procedure. For a 40 000× image, on which a resolution of 10 Å at the specimen is required, a step size of 10 μm [= $(1/4)(10\ \text{Å} \times 40\,000/10\,000\ \text{Å}\ \mu\text{m}^{-1})$] would be suitable.

The best area of an image of a helical or 2D crystal specimen can then be boxed off using a soft-edged mask. For images of single particles, a small, often circular region surrounding each particle is selected to create a stack of many individual particles. In the later steps of image processing, because the orientation and position of each particle are refined by comparing the amplitudes and phases of their Fourier components, it is important to remove spurious features around the edge of each particle and to make sure the different particle images are on the same scale. This is normally done by masking off a circular area centred on each particle and floating the density so that the average around the perimeter becomes zero (DeRosier & Moore, 1970). The edge of the mask is apodized by applying a soft cosine bell shape to the original densities so they taper towards the background level. Finally, to compensate for variations in the exposure due to ice thickness or electron dose, most people normalize the stack of individual particle images so that the mean density and mean density variation over the field of view are set to the same values for all particles (Carrascosa & Steven, 1978).

Once some good particles or crystalline areas for 1D or 2D crystals have been selected, digitized and masked and their intensity values have been normalized, true image processing can begin.

19.6.5. Image processing and 3D reconstruction

Although the general concepts of signal averaging, together with combining different views to reconstruct the 3D structure, are common to the different computer-based procedures that have been implemented, it is important to emphasize one or two preliminary points. First, a homogeneous set of particles must be selected for inclusion in the 3D reconstruction. This selection may be made by eye to eliminate obviously damaged particles or impurities, or by the use of multivariate statistical analysis (van Heel & Frank, 1981) or some other classification scheme, such as local cross-correlation (Roseman, 2003). This allows a subset of the particle images to be used to determine the structure of a better-defined entity. All image-processing procedures require the determination of the same parameters that are needed to specify unambiguously how to combine the information from each micrograph or particle. These parameters are: the magnification, defocus, astigmatism and, at high resolution, the beam tilt for each micrograph; the electron wavelength used (*i.e.* accelerating voltage of the microscope); the spherical aberration coefficient, C_s, of the objective lens; and the orientation and phase origin for each particle or unit cell of the 1D, 2D or 3D crystal. There are 13 parameters for each particle, eight of which may be common to each micrograph and two or three (C_s, accelerating voltage, magnification) to each microscope. The different general approaches that have been used in practice to determine the 3D structure of different classes of macromolecular assemblies from one or more electron micrographs are listed in Table 19.6.5.1.

The precise way in which each general approach codes and determines the particle or unit-cell parameters varies greatly and is not described in detail. Much of the computer software used in image-reconstruction studies is relatively specialized compared with that used in the more mature field of macromolecular X-ray crystallography. In part, this may be attributed to the large diversity of specimen types amenable to cryoTEM and reconstruction methods. As a consequence, image-reconstruction software is still evolving quite rapidly, and references to software packages cited in Table 19.6.5.1 are likely to become quickly outdated. Extensive discussion of algorithms and software packages may be found in several special issues of *Journal of Structural Biology* [Vol. 116, pp. 1–249 (1996); Vol. 120, pp. 207–396 (1997); Vol. 125, pp. 95–254 (1999); Vol. 133, pp. 89–266 (2001); Vol. 138, pp. 1–155 (2002); Vol. 144, pp. 1–252 (2003); Vol. 145, pp. 1–180 (2004); Vol. 157, pp. 1–296 (2007); Vol. 160, pp. 263-404 (2007); and Vol. 161, pp. 219–480 (2008)]. Tomography is the topic of Volumes 120, 138 and 161, single-particle cryoTEM is covered in Volume 133, particle-picking algorithms are covered in Volume 145 and electron crystallography of membrane proteins is covered in Volume 160.

In practice, attempts to determine or refine some parameters may be affected by the inability to determine accurately one of the other parameters. The solution of the structure is therefore an iterative procedure in which reliable knowledge of the parameters that describe each image is gradually built up to produce a more and more accurate structure until no more information can be squeezed out of the micrographs. At this point, if any of the origins or orientations are wrongly assigned, there will be a loss of detail and a decrease in signal-to-noise ratio in the map. If a better-determined or higher-resolution structure is required, it would then be necessary to record images on a better microscope or to prepare new specimens and record better pictures.

The reliability and resolution of the final reconstruction can be measured using a variety of indices. For example, the DPR (Frank *et al.*, 1981), the FSC (van Heel, 1987*b*) and the Q factor (van Heel & Hollenberg, 1980) are three such measures. The DPR is the mean phase difference, as a function of resolution, between the structure factors from two independent reconstructions, often calculated by splitting the image data into two halves. The FSC is a similar calculation of the mean correlation

Table 19.6.5.1
Methods of three-dimensional image reconstruction

Structure type (symmetry)	Method	Reference(s) to technical/theoretical details
Asymmetric (point group C_1)	Random conical tilt	Radermacher *et al.* (1987); Radermacher (1988)
	(software packages)	Frank *et al.* (1996); Baxter *et al.* (2007); Shaikh *et al.* (2008)
	Angular reconstitution	van Heel (1987*a*); Schatz *et al.* (1995)
	(software package)	van Heel *et al.* (1996)
	Weighted back-projection	Radermacher (1991, 1992)
	Radon transform alignment	Radermacher (1994); Ludtke *et al.* (1999); Tang *et al.* (2007)
	Reference-based alignment	Penczek *et al.* (1994)
	Reference-free alignment	Schatz & van Heel (1990); Penczek *et al.* (1992)
	Simulated-annealing alignment	Ogura & Sato (2006)
	Cross-correlation of common lines	Hall *et al.* (2007)
	Ab initio maximum likelihood	Prust *et al.* (2009)
	Maximum-likelihood multi-reference refinement	Scheres *et al.* (2005)
	Fourier reconstruction and alignment	Grigorieff (1998, 2007)
	Orthogonal tilt reconstruction	Leschziner & Nogales (2006)
	Tomographic tilt series and remote control of microscope†	Olins *et al.* (1983); Fung *et al.* (1996); Ziese *et al.* (2002, 2003); Zheng *et al.* (2004); Mastronarde (2005); Nickell *et al.* (2005); Zheng *et al.* (2007)
Symmetric (point groups C_n, D_n; $n > 1$)	Angular reconstitution	van Heel (1987*a*); Schatz *et al.* (1995)
	(software package)	van Heel *et al.* (1996)
	Fourier–Bessel synthesis	Tao *et al.* (1998)
	Reference-based alignment and weighted back-projection	Beuron *et al.* (1998); Ludtke *et al.* (1999); Grigorieff (2007); Tang *et al.* (2007)
Icosahedral (point group I)	Fourier–Bessel synthesis (common lines)	Crowther (1971); Fuller *et al.* (1996); Mancini *et al.* (1997); Thuman-Commike & Chiu (1997)
	Reference-based alignment	Crowther *et al.* (1994); Baker & Cheng (1996); Castón *et al.* (1999); Grigorieff (2007); Yan, Dryden *et al.* (2007)
	(software packages)	Crowther *et al.* (1996); Lawton & Prasad (1996); Zhou *et al.* (1998); Jiang, Li *et al.* (2001); Liang *et al.* (2002); Yan, Sinkovits & Baker (2007)
	Angular reconsitution	van Heel (1987*a*); Stewart *et al.* (1997)
	Model-free reconstruction	Cantele *et al.* (2003); Navaza (2003)
	Multi-path simulated annealing	Liu *et al.* (2007)
	Symmetry-adapted spherical harmonics	Liu, Cheng *et al.* (2008)
	Tomographic tilt series	Walz *et al.* (1997)
Helical	Fourier–Bessel synthesis	DeRosier & Klug (1968); DeRosier & Moore (1970); Stewart (1988); Toyoshima & Unwin (1990); Morgan & DeRosier (1992); Unwin (1993); Beroukhim & Unwin (1997); Miyazawa *et al.* (1999); Wang & Nogales (2005)
	(software packages and filament straightening routines)	Egelman (1986); Whittaker *et al.* (1995); Carragher *et al.* (1996); Crowther *et al.* (1996); Owen *et al.* (1996); Beroukhim & Unwin (1997); Yonekura, Toyoshima *et al.* (2003); Metlagel *et al.* (2007)
	Iterative helical real-space reconstruction	Egelman (2000, 2007*a*); Ramey *et al.* (2009)
2D crystal	Random azimuthal tilt	Henderson & Unwin (1975); Amos *et al.* (1982); Henderson *et al.* (1986); Baldwin *et al.* (1988); Henderson *et al.* (1990)
	(software packages)	Crowther *et al.* (1996); Hardt *et al.* (1996); Gipson, Zeng & Stahlberg (2007); Gipson, Zeng, Zhang & Stahlberg (2007)
	Single-particle refinement	Koeck *et al.* (2007)
3D crystal	Oblique section reconstruction	Crowther & Luther (1984); Winkler & Taylor (1994); Taylor *et al.* (1997)
	(software package)	Winkler & Taylor (1996)
	Sectioned 3D crystal	Winkelmann *et al.* (1991)

† Electron tomography is the subject of several recent reviews (Jonic *et al.*, 2008; McEwen *et al.*, 2008; Morris & Jensen, 2008; Steven & Baumeister, 2008; Bárcena & Koster, 2009; Bartesaghi & Subramaniam, 2009; Hoenger & McIntosh, 2009; Leis *et al.*, 2009; Li & Jensen, 2009), three special issues of *Journal of Structural Biology* [Vol. 120, pp. 207–395 (1997); Vol. 138, pp. 1–155 (2002); and Vol. 161, pp. 219–480 (2008)] and a book edited by Frank (2006*a*).

coefficient between the complex structure factors of the two halves of the data as a function of resolution. The Q factor is the mean ratio of the vector sum of the individual structure factors from each image divided by the sum of their moduli, again calculated as a function of resolution. Perfectly accurate measurements would have values of the DPR, FSC and Q factor of 0°, 1.0 and 1.0, respectively, whereas random data containing no information would have values of 90°, 0.0 and 0.0. The spectral signal-to-noise ratio (SSNR) criterion has been advocated as the best of all (Unser *et al.*, 1989): it effectively measures, as a function of resolution, the overall signal-to-noise ratio (squared) of the whole of the image data. It is calculated by taking into consideration how well all the contributing image data agree internally.

An example of a strategy for determination of the 3D structure of a new and unknown molecule without any symmetry and which does not crystallize might be as follows:

1. Record a single-axis tilt series of particles embedded in negative stain, with a tilt range from −60° to +60°.
2. Calculate 3D structures for each of several different particles using an *R*-weighted back-projection algorithm (Radermacher, 1992).
3. Average 3D data for several particles in real or reciprocal space to get a reasonably good 3D model of the stain-excluding region of the particle.
4. Record a number of micrographs of the particles embedded in vitreous ice.
5. Use the 3D negative-stain model obtained in (3) with inverted contrast to determine the rough alignment parameters of the particle in the ice images.
6. Calculate a preliminary 3D model of the average ice-embedded structure.
7. Use the preliminary 3D model to determine more accurate alignment parameters for the particles in the ice images.
8. Calculate a better 3D model.
9. Determine defocus and astigmatism for each micrograph to allow CTF calculation and correct 3D model so that it represents the structure at high resolution.
10. Keep adding images at different defocus levels to get an accurate structure at as high a resolution as possible.
11. Correct the final structure for resolution-dependent contrast loss and apply a damping factor dependent on signal-to-noise ratio (Rosenthal & Henderson, 2003).

For large single particles with no symmetry, particles with higher symmetry or crystalline arrays, it should be possible to omit the negative-staining steps (1)–(3) and go straight to alignment of particle images from ice embedding, because the particle or crystal tilt angles can be determined internally from comparison of phases along common lines in reciprocal space or from the lattice or helix parameters from a 2D or 1D crystal.

The following discussion briefly outlines for a few specific classes of macromolecule the general strategy for carrying out image processing and 3D reconstruction.

19.6.5.1. 2D crystals

For 2D crystals, the general 3D reconstruction approach consists of the following steps. First, a series of micrographs of single 2D crystals are recorded at different tilt angles, with random azimuthal orientations. Each crystal is then unbent using cross-correlation techniques to identify the precise position of each unit cell (Henderson *et al.*, 1986), and amplitudes and phases of the Fourier components of the average of that particular view of the structure are obtained for the transform of the unbent crystal. The reference image used in the cross-correlation calculation can either be a part of the whole image masked off after a preliminary round of averaging by reciprocal-space filtering of the regions surrounding the diffraction spots in the transform, or it can be a reference image calculated from a previously determined 3D model. The amplitudes and phases from each image are then corrected for the CTF and beam tilt (Henderson *et al.*, 1986, 1990; Kunji *et al.*, 2000) and merged with data from many other crystals by scaling and origin refinement, taking into account the proper symmetry of the 2D space group of the crystal (Landsberg & Hankamer, 2007). Finally, the whole data set is fitted by least squares to constrained amplitudes and phases along the lattice lines (Agard, 1983) prior to calculating a 3D map of the structure. The initial determination of the 2D space group can be carried out by a statistical test of the phase relationships in one or two images of untilted specimens (Valpuesta *et al.*, 1994). The absolute hand of the structure is automatically correct since the 3D structure is calculated from images whose tilt axes and tilt angles are known. Nevertheless, care must be taken not to make any of a number of trivial mistakes that would invert the hand. Recent efforts have been devoted to create a user-friendly interface for image processing of 2D crystals (Gipson, Zeng & Stahlberg, 2007; Gipson, Zeng, Zhang & Stahlberg, 2007; Zeng *et al.*, 2007). The entire topic of processing images of 2D crystals and evaluating the data has been the subject of several publications and reviews (*e.g.* Yeager *et al.*, 1999; Unger, 2000; Renault *et al.*, 2006; Hite *et al.*, 2007; Kovacs *et al.*, 2008).

19.6.5.2. Helical particles

The basic steps involved in processing and 3D reconstruction of helical specimens include the following. First, record a series of micrographs of vitrified particles suspended over holes in a perforated carbon support film. The micrographs are digitized and Fourier transformed to determine image quality (astigmatism, drift, defocus, presence and quality of layer lines, *etc.*). Individual particle images are boxed, floated and apodized within a rectangular mask. The parameters of helical symmetry (number of subunits per turn and pitch) must be determined by indexing the computed diffraction patterns. If necessary, simple spline-fitting procedures may be employed to 'straighten' images of curved particles (Egelman, 1986), and the image data may be reinterpolated (Owen *et al.*, 1996) to provide more precise sampling of the layer-line data in the computed transform. Once a preliminary 3D structure is available, a much more sophisticated refinement of all the helical parameters can be used to unbend the helices onto a predetermined average helix so that the contributions of all parts of the image are correctly treated (Beroukhim & Unwin, 1997). The layer-line data are extracted from each particle transform and two phase-origin corrections are made: one to shift the phase origin to the helix axis (at the centre of the particle image) and the other to correct for effects caused by having the helix axis tilted out of the plane normal to the electron beam in the electron microscope. The layer-line data are separated out into near- and far-side data, corresponding to contributions from the near and far sides of each particle imaged. The relative rotations and translations needed to align the different transforms are determined so the data may be merged and a 3D reconstruction computed by Fourier–Bessel inversion procedures (DeRosier & Moore, 1970). Determination of the absolute hand requires comparison of a pair of images recorded with a small tilt of the specimen between the views (Finch, 1972). An iterative, real-space helical reconstruction method that uses single-particle processing procedures and offers an alternative to the Fourier–Bessel strategy has recently been developed and provides clear advantages for studying structures with imperfect helical symmetry (Egelman 2000, 2007*a*,*b*). Back-projection approaches offer a robust way of dealing with irregular helices or those with discontinuities or seams (Sosa & Milligan, 1996; Sosa *et al.*, 1997).

19.6.5.3. Icosahedral particles

The typical strategy for processing and 3D reconstruction of icosahedral particles consists of the following steps. First, a series of micrographs of a monodisperse distribution of particles, normally suspended over holes in a perforated carbon support film, is recorded. After digitization of the micrographs, individual particle images are boxed and floated with a circular mask. The astigmatism and defocus of each micrograph is measured from the sum of intensities of the Fourier transforms of all particle images (Zhou *et al.*, 1996; Mallick *et al.*, 2005), or by a more automatic procedure applied to the whole micrograph (Mindell & Grigorieff, 2003). Auto-correlation techniques are then used to estimate the particle phase origins, which coincide with the centre of each particle where all rotational symmetry axes intersect (Olson & Baker, 1989). The view orientation of each particle, defined by three Eulerian angles, was originally determined by means of common and cross-common lines techniques (Crowther *et al.*, 1970; Crowther, 1971; Fuller *et al.*, 1996; Baker *et al.*, 1999; Thuman-Commike & Chiu, 2000), but is now more commonly determined with the aid of model-based, template-matching procedures (*e.g.* Baker & Cheng, 1996; Castón *et al.*, 1999; Grigorieff, 2007; Yan, Dryden *et al.*, 2007). Once a set of self-consistent particle images is available, an initial low-resolution 3D reconstruction is computed by merging these data with Fourier–Bessel (Crowther, 1971) or Fourier–Cartesian (*e.g.* Ji *et al.*, 2006) methods. This reconstruction then serves as a reference for refining the orientation, origin and CTF parameters of each of the included particle images, for rejecting 'bad' images, and for increasing the size of the data set by including new particle images from additional micrographs taken at different defocus levels. A new reconstruction, computed from the latest set of images, serves as a new reference and the above refinement procedure is repeated until no further improvements as measured by the reliability criteria mentioned above are made. Determination of the absolute hand of the structure requires the recording and processing of a pair of images taken with a known, small relative tilt of the specimen between the two views (Belnap *et al.*, 1997; Rosenthal & Henderson, 2003). In the past decade, 3D cryoTEM analysis of icosahedral viruses at sub-nanometre resolution has fully matured and near-atomic level (<4 Å) maps (*e.g.* Yu, Jin & Zhou, 2008; Zhang *et al.*, 2008; Zhou, 2008; Chen *et al.*, 2009) will certainly become more commonplace.

19.6.5.4. Electron cryo-tomography

This section is probably more appropriate to a cell-biology textbook than to *International Tables for Crystallography*, but we include it because of its growing importance and potential in the field of structural biology. Good reviews are to be found in Frank (2006*a*) and in three special issues of *Journal of Structural Biology* already mentioned in Table 19.6.5.1. Briefly, a tilt series of micrographs of the specimen is recorded, typically through an angular range of ±60° or ±70°, and this is used, by *R*-weighted back projection, to produce a 3D map of the specimen. This can be further processed, segmented, coloured, rendered and interpreted (*e.g.* Medalia *et al.*, 2002; Nicastro *et al.*, 2005).

Electron cryo-tomography is typically used in two different ways. First, it is used to describe the structure of a unique, non-repeating biological sample such as a cell or organelle, possibly to observe how the structure can change or to explore the diversity of structures that occur. Second, it is used to produce a 3D map (tomogram) of a structure that includes many copies of identical assemblies, such as the trimeric glycoprotein surface spikes on HIV (*e.g.* Liu, Bartesaghi *et al.*, 2008). Individual 3D volumes are then selected from a number of tomograms to give several hundred or thousand sub-tomogram volumes, each of which contains a single copy of the object of interest viewed in an arbitrary direction. By averaging appropriately aligned 3D volumes, which represent a variety of orientations of presumed, identical structures, it is then possible to carry out the same type of averaging as described for all the other specimen types. Although the resolutions attained in the raw tomograms at present (~2009) range at best between 4 and 5 nm, and signal-to-noise considerations (Rosenthal & Henderson, 2003) suggest that the resolution of raw electron cryo-tomograms can never reach better than 1.5–2.0 nm, the procedure of sub-tomogram averaging of identical structures or substructures can, in principle, improve the resolution greatly. This second application of electron cryo-tomography thus fits perfectly into the context of *International Tables*. Examples of this application include the works by Fernandez *et al.* (2006), Zanetti *et al.* (2006), Liu, Bartesaghi *et al.* (2008), Zhu *et al.* (2008) and Zanetti *et al.* (2009). In the area of cellular tomography, the use of high-pressure freezing and cryo-sectioning followed by electron cryo-tomography is most promising (Zuber *et al.*, 2005, 2008; Salje *et al.*, 2009).

19.6.6. Visualization, modelling and interpretation of results

The explosive growth in the number of structures of macromolecules and complexes that are now being determined at sub-nanometre resolutions has spawned the development of numerous tools that enable investigators to visualize, model and interpret their results. Generally, after assessing a global view of the gross morphology of a molecule, one quickly zooms in to explore finer and finer details. A major challenge, especially with density maps of increasing size (*e.g.* >512 pixels), is to perform these operations in real time. Several computer-graphics applications have emerged that satisfy a variety of requirements for visualizing cryoTEM maps. Examples of specific programs include *Chimera* (Pettersen *et al.*, 2004; Goddard *et al.*, 2005; Goddard & Ferrin, 2007; Goddard *et al.*, 2007), *colorRNA* (LeBarron *et al.*, 2007), *MODELLER* (Eswar *et al.*, 2007), *Python Molecular Viewer* (*PMV*) (Sanner, 1999; Gillet *et al.*, 2005), *Vision* (Sanner, 2005); *VMD* (Hsin *et al.*, 2008) and *VolRover* (Bajaj *et al.*, 2005). Many of these programs as well as most of the more common, multipurpose reconstruction programs provide tools for generating animations that can be used to help simulate data and that are particularly useful in communicating results to scientific as well as lay audiences (McGill, 2008). Additional useful tools allow one to produce space-filling, cartoon-like renderings of molecules that reduce unnecessary detail and hence are more easily understood (Goodsell, 2005).

Visualization tools merely provide an entryway to more careful scrutiny and interpretation of structure. When viewed in the context of existing biochemical and molecular studies as well as complementary X-ray crystallographic and other biophysical measurements, structural results provide a basis for understanding how molecules function. Even in the absence of complementary, high-resolution structural data, since 3D cryoTEM maps often reach sub-nanometre resolutions, it is usually possible to recognize characteristic, secondary structural elements like α-helices and large β-sheets in proteins. Several programs help identify such elements (Jiang, Baker *et al.*, 2001; Kong & Ma, 2003; Kong *et al.*, 2004; Baker *et al.*, 2007), and this then facilitates the process of defining and segmenting molecular envelopes (*e.g.* Volkmann, 2002; Baker, Yu *et al.*, 2006) and elucidating protein folds (Zhou *et al.*, 2001; Chiu *et al.*, 2002; Kovacs *et al.*, 2007; Lindert, Staritzbichler *et al.*, 2009). When

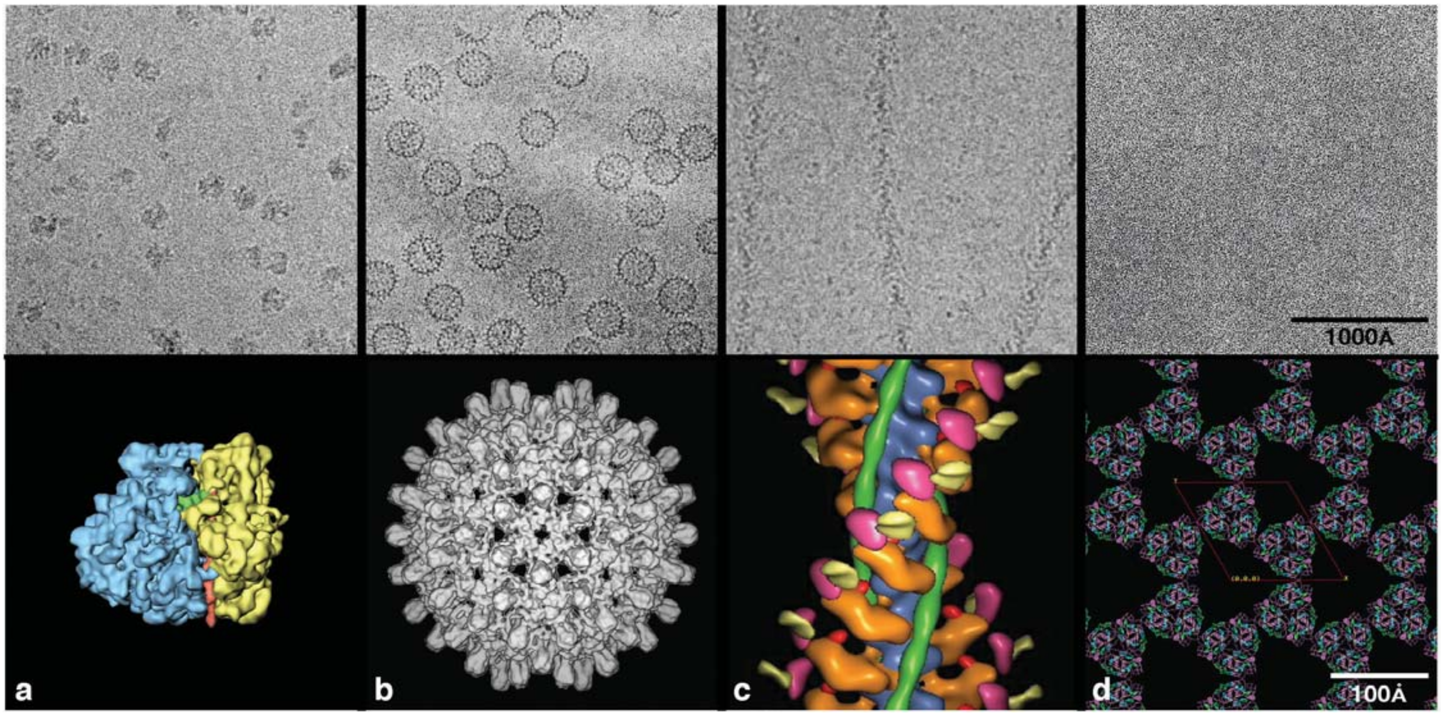


Figure 19.6.7.1
Examples of macromolecules studied by cryoTEM and 3D image reconstruction and the resulting 3D structures (bottom row) after cryoTEM analysis. All micrographs (top row) are displayed at ~170 000× magnification and all models at ~1 200 000× magnification. (*a*) A single particle without symmetry. The micrograph shows 70S *E. coli* ribosomes complexed with mRNA and fMet-tRNA. The surface-shaded density map, made by averaging 73 000 ribosome images from 287 micrographs, has a resolution of 11.5 Å. The 50S and 30S subunits and the tRNA are coloured blue, yellow and green, respectively. The identity of many of the protein and RNA components is known and some RNA double helices are clearly recognizable by their major and minor grooves (*e.g.* helix 44 is shown in red). Courtesy of J. Frank, using data from Gabashvili *et al.* (2000). (*b*) A single particle with symmetry. The micrograph shows hepatitis B virus cores. The 3D reconstruction, at a resolution of 7.4 Å, was computed from 6384 particle images taken from 34 micrographs. From Böttcher *et al.* (1997). (*c*) A helical filament. The micrograph shows actin filaments decorated with myosin S1 heads containing the essential light chain. The 3D reconstruction, at a resolution of 30–35 Å, is a composite in which the differently coloured parts are derived from a series of difference maps that were superimposed on F-actin. The components include: F-actin (blue), myosin heavy-chain motor domain (orange), essential light chain (purple), regulatory light chain (yellow), tropomyosin (green) and myosin motor domain N-terminal beta-barrel (magenta). Courtesy of A. Lin, M. Whittaker & R. Milligan (Scripps Research Institute, La Jolla). (*d*) A 2D crystal: light-harvesting complex LHCII at 3.4 Å resolution. The model shows the protein backbone and the arrangement of chromophores in a number of trimeric subunits in the crystal lattice. In this example, image contrast is too low to see any hint of the structure without image processing (see also Fig. 19.6.4.2). Courtesy of W. Kühlbrandt (Max-Planck-Institute for Biophysics, Frankfurt).

complementary data from X-ray crystallography are available for individual components of a macromolecule, or perhaps even the entire complex, so-called pseudo-atomic models can be produced that closely match the cryoTEM map and lead to more accurate and reliable data interpretation.

Pseudo-atomic models were originally constructed by docking X-ray-derived models manually into cryoTEM density maps (Wang *et al.*, 1992; Rayment *et al.*, 1993; Stewart *et al.*, 1993). However, qualitative procedures like these quickly spurred the development of rigorous, quantitative tools for fitting models into 3D maps as rigid bodies (Volkmann & Hanein, 1999; Wriggers *et al.*, 1999; Roseman, 2000; Rossmann *et al.*, 2001; Wriggers & Chacon, 2001; Chacón & Wriggers, 2002; Navaza *et al.*, 2002; Birmanns & Wriggers, 2003) and several of these methods have been the subject of review articles (Volkmann & Hanein, 2003; Fabiola & Chapman, 2005). The limitations inherent in assuming that vitrified molecules retain the same conformation found in crystals spawned the evolution of numerous flexible fitting and refinement methods (*e.g.* Wriggers & Birmanns, 2001; Gao & Frank, 2005; Velazquez-Muriel *et al.*, 2006; Tan *et al.*, 2008; Dimaio *et al.*, 2009). Many of the flexible fitting methods are based on normal-mode analysis (Tama *et al.*, 2003, 2004*a*,*b*; Bahar & Rader, 2005; Suhre *et al.*, 2006; Gorba *et al.*, 2008) and on molecular and damped dynamics (Kovacs *et al.*, 2008; Orzechowski & Tama, 2008; Trabuco *et al.*, 2008, 2009). When no X-ray structure is available to guide pseudo-atomic modelling of the cryoTEM map, it is still feasible to apply hybrid approaches (*e.g.* iterative structure-based sequence alignment, fitting and refinement) and construct a series of homology models (Topf *et al.*, 2005; Baker, Jiang *et al.*, 2006; Topf *et al.*, 2006; Lindert, Stewart & Meiler, 2009) that can be assessed to help identify reliable candidates (Shacham *et al.*, 2007; Velazquez-Muriel & Carazo, 2007).

19.6.7. Trends

The new generation of intermediate-voltage (~300 kV) FEG microscopes with stable cold stages is now making it much easier to obtain higher-resolution images that, by use of larger defocus values, have good image contrast at both very low and very high resolution. The greater contrast at low resolution greatly facilitates particle-alignment procedures, and the increased contrast resulting from the high-coherence illumination helps to increase the signal-to-noise ratio for the structure at high resolution. Cold stages are constantly being improved, with several liquid-helium stages developed in the 1990s (Fujiyoshi *et al.*, 1991; Zemlin *et al.*, 1996). Two such stages are commercially available from JEOL and FEI.

Finally, some likely trends in the next few years include the following.

(1) Increased automation including:
  (*a*) specimen loading (Lefman *et al.*, 2007);
  (*b*) the recording of micrographs (Potter *et al.*, 1999; Zhang *et al.*, 2003; Lei & Frank, 2005; Suloway *et al.*, 2005; Zheng *et al.*, 2006; Nickell *et al.*, 2007; Yoshioka *et al.*, 2007; Shi *et al.*, 2008; Suloway *et al.*, 2009; Zhang *et al.*, 2009);
  (*c*) particle picking (Short, 2004; Zhu *et al.*, 2004; Chen & Grigorieff, 2007; Plaisier *et al.*, 2007; Woolford *et al.*, 2007; Sorzano *et al.*, 2009; Voss *et al.*, 2009), CTF estimation (*e.g.* Huang *et al.*, 2003; Fernando & Fuller, 2007; Jonic *et al.*, 2007; Sorzano *et al.*, 2007);
  (*d*) 3D reconstructions (*e.g.* Jiang, Li *et al.*, 2001; Mouche *et al.*, 2003; Stagg *et al.*, 2006; Yan, Sinkovits & Baker, 2007; Lander *et al.*, 2009); and
  (*e*) high-performance computing (*e.g.* Bilbao-Castro *et al.*, 2006; Castano-Diez *et al.*, 2007; Yang *et al.*, 2007; Castano-Diez *et al.*, 2008; Fernandez, 2008; Bilbao-Castro *et al.*, 2009).

(2) Introduction of better electronic cameras including:
  (*a*) lens-coupled CCDs (*e.g.* Maki-Yonekura & Yonekura, 2008);
  (*b*) complementary metal-oxide semiconductor (CMOS) sensors, which detect electrons directly. These are based on a technology used in many digital cameras and are displacing CCD technology for larger area detectors (Faruqi *et al.*, 2005; Milazzo *et al.*, 2005; Turchetta *et al.*, 2006; Xuong *et al.*, 2007; Jin *et al.*, 2008);
  (*c*) back-thinned CMOS sensors (McMullan, Faruqi *et al.*, 2009); and
  (*d*) silicon pixel detectors (*e.g.* Faruqi *et al.*, 2003; McMullan *et al.*, 2007).

(3) Increased use of dose-fractionated, tomographic tilt series to extend EM studies to the domain of larger supramolecular and cellular structures (*e.g.* McEwen *et al.*, 1995; Medalia *et al.*, 2002; Lucic *et al.*, 2005; Nicastro *et al.*, 2005, 2006; Jensen & Briegel, 2007; Zheng *et al.*, 2007; Salje *et al.*, 2009).

(4) Evaluation of C_c correctors for cryo-tomography of thicker specimens (Kabius *et al.*, 2009).

(5) The development and use of phase plates (*e.g.* Danev & Nagayama, 2001; Majorovits *et al.*, 2007; Danev & Nagayama, 2008; Nagayama & Danev, 2008; Danev *et al.*, 2009).

We are greatly indebted to all our colleagues in San Diego and Cambridge for their insightful comments and suggestions, to B. Böttcher, R. Crowther, J. Frank, W. Kühlbrandt and R. Milligan for supplying images used in Fig. 19.6.7.1, which gives some examples of the best work done recently, and to N. Olson for identifying numerous recent citations. TSB was supported in part by grant R37 GM-033050 from the National Institutes of Health.

References

Adrian, M., Dubochet, J., Lepault, J. & McDowall, A. W. (1984). *Cryo-electron microscopy of viruses. Nature (London)*, **308**, 32–36.

Agard, D. A. (1983). *A least-squares method for determining structure factors in three-dimensional tilted-view reconstructions. J. Mol. Biol.* **167**, 849–852.

Agirrezabala, X., Martin-Benito, J., Valle, M., Gonzalez, J. M., Valencia, A., Valpuesta, J. M. & Carrascosa, J. L. (2005). *Structure of the connector of bacteriophage T7 at 8 Å resolution: structural homologies of a basic component of a DNA translocating machinery. J. Mol. Biol.* **347**, 895–902.

Amos, L. A., Henderson, R. & Unwin, P. N. T. (1982). *Three-dimensional structure determination by electron microscopy of two-dimensional crystals. Prog. Biophys. Mol. Biol.* **39**, 183–231.

Bahar, I. & Rader, A. J. (2005). *Coarse-grained normal mode analysis in structural biology. Curr. Opin. Struct. Biol.* **15**, 586–592.

Bajaj, C., Castrillon-Candas, J., Siddavanahalli, V. & Xu, Z. (2005). *Compressed representations of macromolecular structures and properties. Structure*, **13** 463–471.

Baker, M. L., Jiang, W., Wedemeyer, W. J., Rixon, F. J., Baker, D. & Chiu, W. (2006) *Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. PLoS Comput. Biol.* **2**, e146.

Baker, M. L., Ju, T. & Chiu, W. (2007). *Identification of secondary structure elements in intermediate-resolution density maps. Structure*, **15**, 7–19.

Baker, M. L., Yu, Z., Chiu, W. & Bajaj, C. (2006). *Automated segmentation of molecular subunits in electron cryomicroscopy density maps. J. Struct. Biol.* **156**, 432–441.

Baker, T. S. & Cheng, R. H. (1996). *A model-based approach for determining orientations of biological macromolecules imaged by cryoelectron microscopy. J. Struct. Biol.* **116**, 120–130.

Baker, T. S., Olson, N. H. & Fuller, S. D. (1999). *Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. Microbiol. Mol. Biol. Rev.* **63**, 862–922.

Baldwin, J. M., Henderson, R., Beckman, E. & Zemlin, F. (1988). *Images of purple membrane at 2.8 Å resolution obtained by cryo-electron microscopy. J. Mol. Biol.* **202**, 585–591.

Bárcena, M. & Koster, A. J. (2009). *Electron tomography in life science. Sem. Cell Develop. Biol.* **20**, 920–930.

Bartesaghi, A. & Subramaniam, S. (2009). *Membrane protein structure determination using cryo-electron tomography and 3D image averaging. Curr. Opin. Struct. Biol.* **19**, 402–407.

Baxter, W. T., Leith, A. & Frank, J. (2007). *SPIRE: the SPIDER reconstruction engine. J. Struct. Biol.* **157**, 56–63.

Bellare, J. R., Davis, H. T., Scriven, L. E. & Talmon, Y. (1988). *Controlled environment vitrification system: an improved sample preparation technique. J. Electron Microsc. Tech.* **10**, 87–111.

Belnap, D. M., Olson, N. H. & Baker, T. S. (1997). *A method for establishing the handedness of biological macromolecules. J. Struct. Biol.* **120**, 44–51.

Beroukhim, R. & Unwin, N. (1997). *Distortion correction of tubular crystals: improvements in the acetylcholine receptor structure. Ultramicroscopy*, **70**, 57–81.

Berriman, J. & Unwin, N. (1994). *Analysis of transient structures by cryo-microscopy combined with rapid mixing of spray droplets. Ultramicroscopy*, **56**, 241–252.

Beuron, F., Maurizi, M. R., Belnap, D. M., Kocsis, E., Booy, F. P., Kessel, M. & Steven, A. C. (1998). *At sixes and sevens: characterization and the symmetry mismatch of the ClpAP chaperone-assisted protease. J. Struct. Biol.* **123**, 248–259.

Bilbao-Castro, J. R., Carazo, J. M., Garcia, I. & Fernandez, J. J. (2006). *Parallelization of reconstruction algorithms in three-dimensional electron microscopy. Appl. Math. Model.* **30**, 688–701.

Bilbao-Castro, J. R., Marabini, R., Sorzano, C. O., Garcia, I., Carazo, J. M. & Fernandez, J. J. (2009). *Exploiting desktop supercomputing for three-dimensional electron microscopy reconstructions using ART with blobs. J. Struct. Biol.* **165**, 19–26.

Birmanns, S. & Wriggers, W. (2003). *Interactive fitting augmented by force-feedback and virtual reality. J. Struct. Biol.* **144**, 123–131.

Bloomer, A. C., Graham, J., Hovmoller, S., Butler, P. J. G. & Klug, A. (1978). *Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. Nature (London)*, **276**, 362–368.

Booth, C. R., Jakana, J. & Chiu, W. (2006). *Assessing the capabilities of a 4k×4k CCD camera for electron cryo-microscopy at 300kV. J. Struct. Biol.* **156**, 556–563.

Booth, C. R., Jiang, W., Baker, M. L., Zhou, Z. H., Ludtke, S. J. & Chiu, W. (2004). *A 9 Å single particle reconstruction from CCD captured images on a 200 kV electron cryomicroscope. J. Struct. Biol.* **147**, 116–127.

Böttcher, B., Wynne, S. A. & Crowther, R. A. (1997). *Determination of the fold of the core protein of hepatitis B virus by electron microscopy. Nature (London)*, **386**, 88–91.

Brenner, S. & Horne, R. W. (1959). *A negative staining method for high resolution electron microscopy of viruses. Biochem. Biophys. Acta Protein Struct.* **34**, 103–110.

Brink, J., Sherman, M. B., Berriman, J. & Chiu, W. (1998). *Evaluation of charging on macromolecules in electron cryomicroscopy. Ultramicroscopy*, **72**, 41–52.

Burmeister, W. P. (2000). *Structural changes in a cryo-cooled protein crystal owing to radiation damage. Acta Cryst.* D**56**, 328–341.

Cantele, F., Lanzavecchia, S. & Bellon, P. L. (2003). *The variance of icosahedral virus models is a key indicator in the structure determination: a model-free reconstruction of viruses, suitable for refractory particles. J. Struct. Biol.* **141**, 84–92.

Carragher, B., Whittaker, M. & Milligan, R. A. (1996). *Helical processing using PHOELIX. J. Struct. Biol.* **116**, 107–112.

Carrascosa, J. L. & Steven, A. C. (1978). *A procedure for evaluation of significant structural differences between related arrays of protein molecules. Micron*, **9**, 199–206.

Castano-Diez, D., Moser, D., Schoenegger, A., Pruggnaller, S. & Frangakis, A. S. (2008). *Performance evaluation of image processing algorithms on the GPU. J. Struct. Biol.* **164**, 153–160.

Castano-Diez, D., Mueller, H. & Frangakis, A. S. (2007). *Implementation and performance evaluation of reconstruction algorithms on graphics processors. J. Struct. Biol.* **157**, 288–295.

Castón, J. R., Belnap, D. M., Steven, A. C. & Trus, B. L. (1999). *A strategy for determining the orientations of refractory particles for reconstruction from cryo-electron micrographs with particular reference to round, smooth-surfaced, icosahedral viruses. J. Struct. Biol.* **125**, 209–215.

Chacón, P. & Wriggers, W. (2002). *Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol.* **317**, 375–384.

Chandramouli, P., Topf, M., Menetret, J. F., Eswar, N., Cannone, J. J., Gutell, R. R., Sali, A. & Akey, C. W. (2008). *Structure of the mammalian 80S ribosome at 8.7 Å resolution. Structure*, **16**, 535–548.

Chen, D. H., Jakana, J., Liu, X., Schmid, M. F. & Chiu, W. (2008). *Achievable resolution from images of biological specimens acquired from a 4k × 4k CCD camera in a 300-kV electron cryomicroscope. J. Struct. Biol.* **163**, 45–52.

Chen, J. Z. & Grigorieff, N. (2007). *SIGNATURE: a single-particle selection system for molecular electron microscopy. J. Struct. Biol.* **157**, 168–173.

Chen, J. Z., Settembre, E. C., Aoki, S. T., Zhang, X., Bellamy, A. R., Dormitzer, P. R., Harrison, S. C. & Grigorieff, N. (2009). *Molecular interactions in rotavirus assembly and uncoating seen by high-resolution cryo-EM. Proc. Natl Acad. Sci. USA*, **106**, 10644–10648.

Cheng, K., Koeck, P. J., Elmlund, H., Idakieva, K., Parvanova, K., Schwarz, H., Ternstrom, T. & Hebert, H. (2006). *Rapana thomasiana hemocyanin (RtH): comparison of the two isoforms, RtH1 and RtH2, at 19 Å and 16 Å resolution. Micron*, **37**, 566–576.

Cheng, R. H., Olson, N. H. & Baker, T. S. (1992). *Cauliflower mosaic virus, a 420 subunit (T = 7), multi-layer structure. Virology*, **186**, 655–668.

Cheng, Y. & Walz, T. (2009). *The advent of near-atomic resolution in single-particle electron microscopy. Annu. Rev. Biochem.* **78**, 723–742.

Cheng, Y., Zak, O., Aisen, P., Harrison, S. C. & Walz, T. (2004). *Structure of the human transferrin receptor–transferrin complex. Cell*, **116**, 565–576.

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M. & Schmid, M. F. (2005). *Electron cryomicroscopy of biological machines at subnanometer resolution. Structure*, **13**, 363–372.

Chiu, W., Baker, M. L., Jiang, W. & Zhou, Z. H. (2002). *Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. Curr. Opin. Struct. Biol.* **12**, 263–269.

Conway, J. F. & Steven, A. C. (1999). *Methods for reconstructing density maps of 'single' particles from cryoelectron micrographs to subnanometer resolution. J. Struct. Biol.* **128**, 106–118.

Craig, L., Volkmann, N., Arvai, A. S., Pique, M. E., Yeager, M., Egelman, E. H. & Tainer, J. A. (2006). *Type IV pilus structure by cryo-electron microscopy and crystallography: implications for pilus assembly and functions. Mol. Cell*, **23**, 651–662.

Crewe, A. V. (1983). *High-resolution scanning transmission electron microscopy. Science*, **221**, 325–330.

Crowther, R. A. (1971). *Procedures for three-dimensional reconstruction of spherical viruses by Fourier synthesis from electron micrographs. Philos. Trans. R. Soc. London*, **261**, 221–230.

Crowther, R. A., Amos, L. A., Finch, J. T., DeRosier, D. J. & Klug, A. (1970). *Three dimensional reconstructions of spherical viruses by Fourier synthesis from electron micrographs. Nature (London)*, **226**, 421–425.

Crowther, R. A., Henderson, R. & Smith, J. M. (1996). *MRC image processing programs. J. Struct. Biol.* **116**, 9–16.

Crowther, R. A., Kiselev, N. A., Böttcher, B., Berriman, J. A., Borisova, G. P., Ose, V. & Pumpens, P. (1994). *Three-dimensional structure of hepatitis B virus core particles determined by electron cryomicroscopy. Cell*, **77**, 943–950.

Crowther, R. A. & Luther, P. K. (1984). *Three-dimensional reconstruction from a single oblique section of fish muscle M-band. Nature (London)*, **307**, 569–570.

Cyrklaff, M. & Kühlbrandt, W. (1994). *High resolution electron microscopy of biological specimens in cubic ice. Ultramicroscopy*, **55**, 141–153.

Danev, R., Glaeser, R. M. & Nagayama, K. (2009). *Practical factors affecting the performance of a thin-film phase plate for transmission electron microscopy. Ultramicroscopy*, **109**, 312–325.

Danev, R. & Nagayama, K. (2001). *Transmission electron microscopy with Zernike phase plate. Ultramicroscopy*, **88**, 243–252.

Danev, R. & Nagayama, K. (2008). *Single particle analysis based on Zernike phase contrast transmission electron microscopy. J. Struct. Biol.* **161**, 211–218.

Dang, T. X., Farah, S. J., Gast, A., Robertson, C., Carragher, B., Egelman, E. & Wilson-Kubalek, E. M. (2005). *Helical crystallization on lipid nanotubes: streptavidin as a model protein. J. Struct. Biol.* **150**, 90–99.

DeRosier, D. J. & Klug, A. (1968). *Reconstruction of three dimensional structures from electron micrographs. Nature (London)*, **217**, 130–134.

DeRosier, D. J. & Moore, P. B. (1970). *Reconstruction of three-dimensional images from electron micrographs of structures with helical symmetry. J. Mol. Biol.* **52**, 355–369.

Dimaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. (2009). *Refinement of protein structures into low-resolution density maps using Rosetta. J. Mol. Biol.* **392**, 181–190.

Dubochet, J., Adrian, M., Chang, J.-J., Homo, J.-C., Lepault, J., McDowall, A. W. & Schultz, P. (1988). *Cryo-electron microscopy of vitrified specimens. Q. Rev. Biophys.* **21**, 129–228.

Dubochet, J., Lepault, J., Freeman, R., Berriman, J. A. & Homo, J.-C. (1982). *Electron microscopy of frozen water and aqueous solutions. J. Microsc.* **128**, 219–237.

Egelman, E. H. (1986). *An algorithm for straightening images of curved filamentous structures. Ultramicroscopy*, **19**, 367–374.

Egelman, E. H. (2000). *A robust algorithm for the reconstruction of helical filaments using single-particle methods. Ultramicroscopy*, **85**, 225–234.

Egelman, E. H. (2007*a*). *The iterative helical real space reconstruction method: surmounting the problems posed by real polymers. J. Struct. Biol.* **157**, 83–94.

Egelman, E. H. (2007*b*). *Single-particle reconstruction from EM images of helical filaments. Curr. Opin. Struct. Biol.* **17**, 556–561.

Elmlund, H., Lundqvist, J. , Al-Karadaghi, S., Hansson, M., Hebert, H. & Lindahl, M. (2008). *A new cryo-EM single-particle ab initio reconstruction method visualizes secondary structure elements in an ATP-fueled AAA+ motor. J. Mol. Biol.* **375**, 934–947.

Erickson, H. P. & Klug, A. (1971). *Measurement and compensation of defocusing and aberrations by Fourier processing of micrographs. Philos. Trans. R. Soc. London Ser. B*, **261**, 105–118.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. & Sali, A. (2007). *Comparative protein structure modeling using MODELLER. Curr. Protoc. Protein Sci.* ch. 2, unit 2.9.

Fabiola, F. & Chapman, M. S. (2005). *Fitting of high-resolution structures into electron microscopy reconstruction images. Structure*, **13**, 389–400.

Faruqi, A. R., Cattermole, D. M. & Raeburn, C. (2003). *Direct electron detection methods in electron microscopy. Nucl. Instrum. Methods Phys. Res. A*, **513**, 317–321.

Faruqi, A. R., Henderson, R., Pryddetch, M., Allport, P. & Evans, A. (2005). *Direct single electron detection with a CMOS detector for electron microscopy. Nucl. Instrum. Methods Phys. Res. A*, **546**, 170–175.

Fernandez, J. J. (2008). *High performance computing in structural determination by electron cryomicroscopy. J. Struct. Biol.* **164**, 1–6.

Fernandez, J. J., Li, S. & Crowther, R. A. (2006). *CTF determination and correction in electron cryotomography. Ultramicroscopy*, **106**, 587–596.

Fernandez, J. J., Luque, D., Caston, J. R. & Carrascosa, J. L. (2008). *Sharpening high resolution information in single particle electron cryomicroscopy*. *J. Struct. Biol.* **164**, 170–175.

Fernandez-Tornero, C., Bottcher, B., Riva, M., Carles, C., Steuerwald, U., Ruigrok, R. W., Sentenac, A., Muller, C. W. & Schoehn, G. (2007). *Insights into transcription initiation and termination from the electron microscopy structure of yeast RNA polymerase III*. *Mol. Cell*, **25**, 813–823.

Fernando, K. V. & Fuller, S. D. (2007). *Determination of astigmatism in TEM images*. *J. Struct. Biol.* **157**, 189–200.

Finch, J. T. (1972). *The hand of the helix of tobacco mosaic virus*. *J. Mol. Biol.* **66**, 291–294.

Fleishman, S. J., Unger, V. M., Yeager, M. & Ben-Tal, N. (2004). *A Calpha model for the transmembrane alpha helices of gap junction intercellular channels*. *Mol. Cell*, **15**, 879–888.

Ford, R. C., Brunkan-LaMontagne, A. L., Collins, R. F., Clarke, B. R., Harris, R., Naismith, J. H. & Whitfield, C. (2009). *Structure-function relationships of the outer membrane translocon Wza investigated by cryo-electron microscopy and mutagenesis*. *J. Struct. Biol.* **166**, 172–182.

Fotin, A., Cheng, Y., Sliz, P., Grigorieff, N., Harrison, S. C., Kirchhausen, T. & Walz, T. (2004). *Molecular model for a complete clathrin lattice from electron cryomicroscopy*. *Nature (London)*, **432**, 573–579.

Frank, J. (2002). *Single-particle imaging of macromolecules by cryo-electron microscopy*. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319.

Frank, J. (2006*a*). Editor. *Electron Tomography: Methods for Three-Dimensional Visualization of Structures in the Cell*. New York: Springer.

Frank, J. (2006*b*). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in their Native State*. Oxford University Press.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. & Leith, A. (1996). *SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields*. *J. Struct. Biol.* **116**, 190–199.

Frank, J., Verschoor, A. & Boublik, M. (1981). *Computer averaging of electron micrographs of 40S ribosomal subunits*. *Science*, **214**, 1353–1355.

Frost, A., Perera, R., Roux, A., Spasov, K., Destaing, O., Egelman, E. H., De Camilli, P. & Unger, V. M. (2008). *Structural basis of membrane invagination by F-BAR domains*. *Cell*, **132**, 807–817.

Fujiyoshi, Y., Mizusaki, T., Morikawa, K., Yamagishi, H., Aoki, Y., Kihara, H. & Harada, Y. (1991). *Development of a superfluid helium stage for high-resolution electron microscopy*. *Ultramicroscopy*, **38**, 241–251.

Fuller, S. D., Berriman, J. A., Butcher, S. J. & Gowen, B. E. (1995). *Low pH induces swiveling of the glycoprotein heterodimers in the Semliki forest virus spike complex*. *Cell*, **81**, 715–725.

Fuller, S. D., Butcher, S. J., Cheng, R. H. & Baker, T. S. (1996). *Three-dimensional reconstruction of icosahedral particles – the uncommon line*. *J. Struct. Biol.* **116**, 48–55.

Fung, J. C., Liu, W., DeRuijter, W. J., Chen, H., Abbey, C. K., Sedat, J. W. & Agard, D. A. (1996). *Toward fully automated high-resolution electron tomography*. *J. Struct. Biol.* **116**, 181–189.

Gabashvili, I. S., Agrawal, R. K., Spahn, C. M. T., Grassucci, R. A., Svergun, D. I., Frank, J. & Penczek, P. (2000). *Solution structure of the E. coli 70S ribosome at 11.5 Å resolution*. *Cell*, **100**, 537–549.

Galkin, V. E., Orlova, A., Cherepanova, O., Lebart, M. C. & Egelman, E. H. (2008). *High-resolution cryo-EM structure of the F-actin–fimbrin/plastin ABD2 complex*. *Proc. Natl Acad. Sci. USA*, **105**, 1494–1498.

Gao, H. & Frank, J. (2005). *Molding atomic structures into intermediate-resolution cryo-EM density maps of ribosomal complexes using real-space refinement*. *Structure*, **13**, 401–406.

Gatsogiannis, C., Moeller, A., Depoix, F., Meissner, U. & Markl, J. (2007). *Nautilus pompilius hemocyanin: 9 Å cryo-EM structure and molecular model reveal the subunit pathway and the interfaces between the 70 functional units*. *J. Mol. Biol.* **374**, 465–486.

Gillet, A., Sanner, M., Stoffler, D. & Olson, A. (2005). *Tangible interfaces for structural molecular biology*. *Structure*, **13**, 483–491.

Gipson, B., Zeng, X. & Stahlberg, H. (2007). *2dx_merge: data management and merging for 2D crystal images*. *J. Struct. Biol.* **160**, 375–384.

Gipson, B., Zeng, X., Zhang, Z. Y. & Stahlberg, H. (2007). *2dx – user-friendly image processing for 2D crystals*. *J. Struct. Biol.* **157**, 64–72.

Glaeser, R., Downing, K., DeRosier, D., Chiu, W. & Frank, J. (2007). *Electron Crystallography of Biological Macromolecules*. Oxford University Press.

Glaeser, R. M. (1971). *Limitations to significant information in biological electron microscopy as a result of radiation damage*. *J. Ultrastruct. Res.* **36**, 466–482.

Glaeser, R. M. (2008*a*). *Cryo-electron microscopy of biological nano-structures*. *Phys. Today*, **61**, 48–54.

Glaeser, R. M. (2008*b*). *Retrospective: radiation damage and its associated 'information limitations'*. *J. Struct. Biol.* **163**, 271–276.

Goddard, T. D. & Ferrin, T. E. (2007). *Visualization software for molecular assemblies*. *Curr. Opin. Struct. Biol.* **17**, 587–595.

Goddard, T. D., Huang, C. C. & Ferrin, T. E. (2005). *Software extensions to UCSF chimera for interactive visualization of large molecular assemblies*. *Structure*, **13**, 473–482.

Goddard, T. D., Huang, C. C. & Ferrin, T. E. (2007). *Visualizing density maps with UCSF Chimera*. *J. Struct. Biol.* **157**, 281–287.

Gonen, T., Cheng, Y., Sliz, P., Hiroaki, Y., Fujiyoshi, Y., Harrison, S. C. & Walz, T. (2005). *Lipid–protein interactions in double-layered two-dimensional AQP0 crystals*. *Nature (London)*, **438**, 633–638.

Goodsell, D. S. (2005). *Visual methods from atoms to cells*. *Structure*, **13**, 347–354.

Gorba, C., Miyashita, O. & Tama, F. (2008). *Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data*. *Biophys. J.* **94**, 1589–1599.

Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R. & Hogle, J. M. (1998). *The crystal structure of Dps, a ferritin homolog that binds and protects DNA*. *Nat. Struct. Biol.* **5**, 294–303.

Grassucci, R. A., Taylor, D. J. & Frank, J. (2007). *Preparation of macromolecular complexes for cryo-electron microscopy*. *Nat. Protoc.* **2**, 3239–3246.

Gregorini, M., Wang, J., Xie, X. S., Milligan, R. A. & Engel, A. (2007). *Three-dimensional reconstruction of bovine brain V-ATPase by cryo-electron microscopy and single particle analysis*. *J. Struct. Biol.* **158**, 445–454.

Grigorieff, N. (1998). *Three-dimensional structure of bovine NADH:ubiquinone oxidoreductase (complex I) at 22 Å in ice*. *J. Mol. Biol.* **277**, 1033–1046.

Grigorieff, N. (2007). *FREALIGN: high-resolution refinement of single particle structures*. *J. Struct. Biol.* **157**, 117–125.

Hall, R. J., Siridechadilok, B. & Nogales, E. (2007). *Cross-correlation of common lines: a novel approach for single-particle reconstruction of a structure containing a flexible domain*. *J. Struct. Biol.* **159**, 474–482.

Hardt, S., Wang, B. & Schmid, M. F. (1996). *A brief description of I.C.E.: the integrated crystallographic environment*. *J. Struct. Biol.* **116**, 68–70.

Hargittai, I. & Hargittai, M. (1988). Editors. *Stereochemical Applications of Gas-Phase Electron Diffraction*. New York: VCH.

Hasler, L., Heymann, J. B., Engel, A., Kistler, J. & Walz, T. (1998). *2D crystallization of membrane proteins: rationales and examples*. *J. Struct. Biol.* **121**, 162–171.

Havelka, W. A., Henderson, R. & Oesterhelt, D. (1995). *Three-dimensional structure of halorhodopsin at 7 Å resolution*. *J. Mol. Biol.* **247**, 726–738.

He, Y. & Olson, R. (2010). *Three-dimensional structure of the detergent-solubilized Vibrio cholerae cytolysin (VCC) heptamer by electron cryomicroscopy*. *J. Struct. Biol.* **169**, 6–13.

Heel, M. van (1987*a*). *Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction*. *Ultramicroscopy*, **21**, 111–124.

Heel, M. van (1987*b*). *Similarity measures between images*. *Ultramicroscopy*, **21**, 95–100.

Heel, M. van & Frank, J. (1981). *Use of multivariate statistics in analyzing the images of biological macromolecules*. *Ultramicroscopy*, **6**, 187–194.

Heel, M. van, Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. & Patwardhan, A. (2000). *Single-particle electron cryo-microscopy: towards atomic resolution*. *Quart. Rev. Biophys.* **33**, 307–369.

Heel, M. van, Harauz, G. & Orlova, E. V. (1996). *A new generation of the IMAGIC image processing system*. *J. Struct. Biol.* **116**, 17–24.

Heel, M. van & Hollenberg, J. (1980). *On the stretching of distorted images of two-dimensional crystals*. In *Electron Microscopy at Molecular Dimensions*, edited by W. Baumeister & W. Vogell, pp. 256–260. Berlin: Springer-Verlag.

Heitkamp, T., Bottcher, B. & Greie, J. C. (2009). *Solution structure of the KdpFABC P-type ATPase from Escherichia coli by electron microscopic single particle analysis*. *J. Struct. Biol.* **166**, 295–302.

Henderson, R. (1995). *The potential and limitations of neutrons, electrons, and X-rays for atomic resolution microscopy of unstained biological molecules. Q. Rev. Biophys.* **28**, 171–193.

Henderson, R. (2004). *Realizing the potential of electron cryo-microscopy. Quart. Rev. Biophys.* **37**, 3–13.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E. & Downing, K. H. (1990). *Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. J. Mol. Biol.* **213**, 899–929.

Henderson, R., Baldwin, J. M., Downing, K. H., Lepault, J. & Zemlin, F. (1986). *Structure of purple membrane from halobacterium: recording, measurement, and evaluation of electron micrographs at 3.5 Å resolution. Ultramicroscopy*, **19**, 147–178.

Henderson, R. & Unwin, P. N. T. (1975). *Three-dimensional model of purple membrane obtained by electron microscopy. Nature (London)*, **257**, 28–32.

Hite, R. K., Raunser, S. & Walz, T. (2007). *Revival of electron crystallography. Curr. Opin. Struct. Biol.* **17**, 389–395.

Hoenger, A. & McIntosh, J. R. (2009). *Probing the macromolecular organization of cells by electron tomography. Curr. Opin. Cell Biol.* **21**, 89–96.

Holm, P. J., Bhakat, P., Jegerschold, C., Gyobu, N., Mitsuoka, K., Fujiyoshi, Y., Morgenstern, R. & Hebert, H. (2006). *Structural basis for detoxification and oxidative stress protection in membranes. J. Mol. Biol.* **360**, 934–945.

Hoppe, W., Langer, R., Knesch, G. & Poppe, C. (1968). *Protein-kristallstrukturanalyse mit elektronenstrahlen. Naturwissenschaften*, **55**, 333–336.

Hsin, J., Arkhipov, A., Yin, Y., Stone, J. E. & Schulten, K. (2008). *Using VMD: an introductory tutorial. Curr. Protoc. Bioinfor.* ch. 5, unit 5.7.

Huang, Z., Baldwin, P. R., Mullapudi, S. & Penczek, P. A. (2003). *Automated determination of parameters describing power spectra of micrograph images in electron microscopy. J. Struct. Biol.* **144**, 79–94.

Huxley, H. E. & Zubay, G. (1960). *Electron microscope observations on the structure of microsomal particles from Escherichia coli. J. Mol. Biol.* **2**, 10–18.

Iancu, C. V., Tivol, W. F., Schooler, J. B., Dias, D. P., Henderson, G. P., Murphy, G. E., Wright, E. R., Li, Z., Yu, Z., Briegel, A., Gan, L., He, Y. & Jensen, G. J. (2006). *Electron cryotomography sample preparation using the Vitrobot. Nat. Protoc.* **1**, 2813–2819.

Isaacson, M., Langmore, J. & Rose, H. (1974). *Determination of the non-localization of the inelastic scattering of electrons by electron microscopy. Optik*, **41**, 92–96.

Jap, B., Zulauf, M., Scheybani, T., Hefti, A., Baumeister, W. & Aebi, U. (1992). *2D crystallization: from art to science. Ultramicroscopy*, **46**, 45–84.

Jegerschold, C., Pawelzik, S. C., Purhonen, P., Bhakat, P., Gheorghe, K. R., Gyobu, N., Mitsuoka, K., Morgenstern, R., Jakobsson, P. J. & Hebert, H. (2008). *Structural basis for induced formation of the inflammatory mediator prostaglandin E2. Proc. Natl Acad. Sci. USA*, **105**, 11110–11115.

Jensen, G. J. & Briegel, A. (2007). *How electron cryotomography is opening a new window onto prokaryotic ultrastructure. Curr. Opin. Struct. Biol.* **17**, 260–267.

Ji, Y., Marinescu, D. C., Zhang, W., Zhang, X., Yan, X. & Baker, T. S. (2006). *A model-based parallel origin and orientation refinement algorithm for cryoTEM and its application to the study of virus structures. J. Struct. Biol.* **154**, 1–19.

Jiang, W., Baker, M. L., Jakana, J., Weigele, P. R., King, J. & Chiu, W. (2008). *Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. Nature (London)*, **451**, 1130–1134.

Jiang, W., Baker, M. L., Ludtke, S. J. & Chiu, W. (2001). *Bridging the information gap: computational tools for intermediate resolution structure interpretation. J. Mol. Biol.* **308**, 1033–1044.

Jiang, W., Li, Z., Zhang, Z., Booth, C. R., Baker, M. L. & Chiu, W. (2001). *Semi-automated icosahedral particle reconstruction at sub-nanometer resolution. J. Struct. Biol.* **136**, 214–225.

Jin, L., Milazzo, A. C., Kleinfelder, S., Li, S., Leblanc, P., Duttweiler, F., Bouwer, J. C., Peltier, S. T., Ellisman, M. H. & Xuong, N. H. (2008). *Applications of direct detection device in transmission electron microscopy. J. Struct. Biol.* **161**, 352–358.

Jonic, S., Sorzano, C. O. & Boisset, N. (2008). *Comparison of single-particle analysis and electron tomography approaches: an overview. J. Microsc.* **232**, 562–579.

Jonic, S., Sorzano, C. O., Cottevieille, M., Larquet, E. & Boisset, N. (2007). *A novel method for improvement of visualization of power spectra for sorting cryo-electron micrographs and their local areas. J. Struct. Biol.* **157**, 156–167.

Kabius, B., Hartel, P., Haider, M., Muller, H., Uhlemann, S., Loebau, U., Zach, J. & Rose, H. (2009). *First application of Cc-corrected imaging for high-resolution and energy-filtered TEM. J. Electron Microsc.* **58**, 147–155.

Kasas, S., Dumas, G., Dietler, G., Catsicas, S. & Adrian, M. (2003). *Vitrification of cryoelectron microscopy specimens revealed by high-speed photographic imaging. J. Microsc.* **211**, 48–53.

Kenney, J., Karsenti, E., Gowen, B. & Fuller, S. D. (1997). *Three-dimensional reconstruction of the mammalian centriole from cryoelectron micrographs: the use of common lines for orientation and alignment. J. Struct. Biol.* **120**, 320–328.

Klug, A. & Berger, J. E. (1964). *An optical method for the analysis of periodicities in electron micrographs, and some observations on the mechanism of negative staining. J. Mol. Biol.* **10**, 565–569.

Klug, A. & DeRosier, D. J. (1966). *Optical filtering of electron micrographs: reconstruction of one sided images. Nature (London)*, **212**, 29–32.

Koeck, P. J., Purhonen, P., Alvang, R., Grundberg, B. & Hebert, H. (2007). *Single particle refinement in electron crystallography: a pilot study. J. Struct. Biol.* **160**, 344–352.

Kong, Y. & Ma, J. (2003). *A structural-informatics approach for mining β-sheets: locating sheets in intermediate-resolution density maps. J. Mol. Biol.* **332**, 399–413.

Kong, Y., Zhang, X., Baker, T. S. & Ma, J. (2004). *A structural-informatics approach for tracing β-sheets: building pseudo-Cα traces for β-strands in intermediate-resolution density maps. J. Mol. Biol.* **339**, 117–130.

Kornberg, R. & Darst, S. A. (1991). *Two dimensional crystals of proteins on liquid layers. Curr. Opin. Struct. Biol.* **1**, 642–646.

Kovacs, J. A., Yeager, M. & Abagyan, R. (2007). *Computational prediction of atomic structures of helical membrane proteins aided by EM maps. Biophys. J.* **93**, 1950–1959.

Kovacs, J. A., Yeager, M. & Abagyan, R. (2008). *Damped-dynamics flexible fitting. Biophys. J.* **95**, 3192–3207.

Krivanek, O. L. & Mooney, P. E. (1993). *Applications of slow-scan CCD cameras in transmission electron microscopy. Ultramicroscopy*, **49**, 95–108.

Kubalek, E. W., LeGrice, S. F. J. & Brown, P. O. (1994). *Two-dimensional crystallization of histidine-tagged, HIV-1 reverse transcriptase promoted by a novel nickel-chelating lipid. J. Struct. Biol.* **113**, 117–123.

Kühlbrandt, W., Wang, D. N. & Fujiyoshi, Y. (1994). *Atomic model of plant light-harvesting complex by electron crystallography. Nature (London)*, **367**, 614–621.

Kunji, E. R., von Gronau, S., Oesterhelt, D. & Henderson, R. (2000). *The three-dimensional structure of halorhodopsin to 5 Å by electron crystallography: a new unbending procedure for two-dimensional crystals by using a global reference structure. Proc. Natl Acad. Sci. USA*, **97**, 4637–4642.

Lander, G. C., Stagg, S. M., Voss, N. R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P. W., Lyumkis, D., Potter, C. S. & Carragher, B. (2009). *Appion: an integrated, database-driven pipeline to facilitate EM image processing. J. Struct. Biol.* **166**, 95–102.

Landsberg, M. J. & Hankamer, B. (2007). *Symmetry: a guide to its application in 2D electron crystallography. J. Struct. Biol.* **160**, 332–343.

Lau, W. C., Baker, L. A. & Rubinstein, J. L. (2008). *Cryo-EM structure of the yeast ATP synthase. J. Mol. Biol.* **382**, 1256–1264.

Lawton, J. A. & Prasad, B. V. V. (1996). *Automated software package for icosahedral virus reconstruction. J. Struct. Biol.* **116**, 209–215.

LeBarron, J., Grassucci, R. A., Shaikh, T. R., Baxter, W. T., Sengupta, J. & Frank, J. (2008). *Exploration of parameters in cryo-EM leading to an improved density map of the E. coli ribosome. J. Struct. Biol.* **164**, 24–32.

LeBarron, J., Mitra, K. & Frank, J. (2007). *Displaying 3D data on RNA secondary structures: coloRNA. J. Struct. Biol.* **157**, 262–270.

Lefman, J., Morrison, R. & Subramaniam, S. (2007). *Automated 100-position specimen loader and image acquisition system for transmission electron microscopy. J. Struct. Biol.* **158**, 318–326.

Lei, J. & Frank, J. (2005). *Automated acquisition of cryo-electron micrographs for single particle reconstruction on an FEI Tecnai electron microscope. J. Struct. Biol.* **150**, 69–80.

Leis, A., Rockel, B., Andrees, L. & Baumeister, W. (2009). *Visualizing cells at the nanoscale. Trends Biochem. Sci.* **34**, 60–70.

Leschziner, A. E. & Nogales, E. (2006). *The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles*. *J. Struct. Biol.* **153**, 284–299.

Li, H., DeRosier, D. J., Nicholson, W. V., Nogales, E. & Downing, K. H. (2002). *Microtubule structure at 8 Å resolution*. *Structure*, **10**, 1317–1328.

Li, Z. & Jensen, G. J. (2009). *Electron cryotomography: a new view into microbial ultrastructure*. *Curr. Opin. Microbiol.* **12**, 333–340.

Liang, Y., Ke, E. Y. & Zhou, Z. H. (2002). *IMIRS: a high-resolution 3D reconstruction package integrated with a relational image database*. *J. Struct. Biol.* **137**, 292–304.

Lindert, S., Staritzbichler, R., Wotzel, N., Karakas, M., Stewart, P. L. & Meiler, J. (2009). *EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps*. *Structure*, **17**, 990–1003.

Lindert, S., Stewart, P. L. & Meiler, J. (2009). *Hybrid approaches: applying computational methods in cryo-electron microscopy*. *Curr. Opin. Struct. Biol.* **19**, 218–225.

Liu, H., Cheng, L., Zeng, S., Cai, C., Zhou, Z. H. & Yang, Q. (2008). *Symmetry-adapted spherical harmonics method for high-resolution 3D single-particle reconstructions*. *J. Struct. Biol.* **161**, 64–73.

Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G. & Subramaniam, S. (2008). *Molecular architecture of native HIV-1 gp120 trimers*. *Nature (London)*, **455**, 109–113.

Liu, J., Wu, S., Reedy, M. C., Winkler, H., Lucaveche, C., Cheng, Y., Reedy, M. K. & Taylor, K. A. (2006). *Electron tomography of swollen rigor fibers of insect flight muscle reveals a short and variably angled S2 domain*. *J. Mol. Biol.* **362**, 844–860.

Liu, X., Jiang, W., Jakana, J. & Chiu, W. (2007). *Averaging tens to hundreds of icosahedral particle images to resolve protein secondary structure elements using a multi-path simulated annealing optimization algorithm*. *J. Struct. Biol.* **160**, 11–27.

Lucic, V., Forster, F. & Baumeister, W. (2005). *Structural studies by electron tomography: from cells to molecules*. *Annu. Rev. Biochem.* **74**, 833–865.

Ludtke, S. J., Baker, M. L., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2008). *De novo backbone trace of GroEL from single particle electron cryomicroscopy*. *Structure*, **16**, 441–448.

Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). *EMAN: semiautomated software for high-resolution single-particle reconstructions*. *J. Struct. Biol.* **128**, 82–97.

Ludtke, S. J., Serysheva, I. I., Hamilton, S. L. & Chiu, W. (2005). *The pore structure of the closed RyR1 channel*. *Structure*, **13**, 1203–1211.

McEwen, B. F., Downing, K. H. & Glaeser, R. M. (1995). *The relevance of dose-fractionation in tomography of radiation-sensitive specimens*. *Ultramicroscopy*, **60**, 357–373.

McEwen, B. F., Renken, C., Marko, M. & Mannella, C. (2008). *Chapter 6: Principles and practice in electron tomography*. *Methods Cell Biol.* **89**, 129–168.

McGill, G. (2008). *Molecular movies … coming to a lecture near you*. *Cell*, **133**, 1127–1132.

McMullan, G., Cattermole, D. M., Chen, S., Henderson, R., Llopart, X., Summerfield, C., Tlustos, L. & Faruqi, A. R. (2007). *Electron imaging with Medipix2 hybrid pixel detector*. *Ultramicroscopy*, **107**, 401–413.

McMullan, G., Chen, S., Henderson, R. & Faruqi, A. R. (2009). *Detective quantum efficiency of electron area detectors in electron microscopy*. *Ultramicroscopy*, **109**, 1126–1143.

McMullan, G., Faruqi, A. R., Henderson, R., Guerrini, N., Turchetta, R., Jacobs, A. & van Hoften, G. (2009). *Experimental observation of the improvement in MTF from backthinning a CMOS direct electron detector*. *Ultramicroscopy*, **109**, 1144–1147.

Majorovits, E., Barton, B., Schultheiss, K., Perez-Willard, F., Gerthsen, D. & Schroder, R. R. (2007). *Optimizing phase contrast in transmission electron microscopy with an electrostatic (Boersch) phase plate*. *Ultramicroscopy*, **107**, 213–226.

Maki-Yonekura, S. & Yonekura, K. (2008). *Electron digital imaging toward high-resolution structure analysis of biological macromolecules*. *Microsc. Microanal.* **14**, 362–369.

Mallick, S. P., Carragher, B., Potter, C. S. & Kriegman, D. J. (2005). *ACE: automated CTF estimation*. *Ultramicroscopy*, **104**, 8–29.

Mancini, E. J., de Haas, F. & Fuller, S. D. (1997). *High-resolution icosahedral reconstruction: fulfilling the promise of cryo-electron microscopy*. *Structure*, **5**, 741–750.

Marlovits, T. C., Kubori, T., Lara-Tejero, M., Thomas, D., Unger, V. M. & Galan, J. E. (2006). *Assembly of the inner rod determines needle length in the type III secretion injectisome*. *Nature (London)*, **441**, 637–640.

Mastronarde, D. N. (2005). *Automated electron microscope tomography using robust prediction of specimen movements*. *J. Struct. Biol.* **152**, 36–51.

Matadeen, R., Hon, W. C., Heath, J. K., Jones, E. Y. & Fuller, S. (2007). *The dynamics of signal triggering in a gp130-receptor complex*. *Structure*, **15**, 441–448.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., de Kok, A. & Hol, W. G. J. (1992). *Atomic structure of the cubic core of the pyruvate dehydrogenase multienzyme complex*. *Science*, **255**, 1544–1550.

Mayer, E. & Astl, G. (1992). *Limits of cryofixation as seen by Fourier transform infrared spectra of metmyoglobin azide and carbonyl hemoglobin in vitrified and freeze concentrated aqueous solution*. *Ultramicroscopy*, **45**, 185–197.

Medalia, O., Weber, I., Frangakis, A. S., Nicastro, D., Gerisch, G. & Baumeister, W. (2002). *Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography*. *Science*, **298**, 1209–1213.

Meents, A., Gutmann, S., Wagner, A. & Schulze-Briese, C. (2010). *Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures*. *Proc. Natl Acad. Sci. USA*, **107**, 1094–1099.

Melanson, L. (2009). *A versatile and affordable plunge freezing instrument for preparing frozen hydrated specimens for cryo transmission electron microscopy (CryoEM)*. *Microsc. Today*, **17**, 14–17.

Meng, X., Wang, G., Viero, C., Wang, Q., Mi, W., Su, X. D., Wagenknecht, T., Williams, A. J., Liu, Z. & Yin, C. C. (2009). *CLIC2-RyR1 interaction and structural characterization by cryo-electron microscopy*. *J. Mol. Biol.* **387**, 320–334.

Metlagel, Z., Kikkawa, Y. S. & Kikkawa, M. (2007). *Ruby-Helix: an implementation of helical image processing based on object-oriented scripting language*. *J. Struct. Biol.* **157**, 95–105.

Mikyas, Y., Makabi, M., Raval-Fernandes, S., Harrington, L., Kickhoefer, V. A., Rome, L. H. & Stewart, P. L. (2004). *Cryoelectron microscopy imaging of recombinant and tissue derived vaults: localization of the MVP N termini and VPARP*. *J. Mol. Biol.* **344**, 91–105.

Milazzo, A. C., Leblanc, P., Duttweiler, F., Jin, L., Bouwer, J. C., Peltier, S., Ellisman, M., Bieser, F., Matis, H. S., Wieman, H., Denes, P., Kleinfelder, S. & Xuong, N. H. (2005). *Active pixel sensor array as a detector for electron microscopy*. *Ultramicroscopy*, **104**, 152–159.

Milligan, R. A. (1996). *Protein–protein interactions in the rigor actomyosin complex*. *Proc. Natl Acad. Sci. USA*, **93**, 21–26.

Mindell, J. A. & Grigorieff, N. (2003). *Accurate determination of local defocus and specimen tilt in electron microscopy*. *J. Struct. Biol.* **142**, 334–347.

Miyazawa, A., Fujiyoshi, Y., Stowell, M. & Unwin, N. (1999). *Nicotinic acetylcholine receptor at 4.6 Å resolution: transverse tunnels in the channel wall*. *J. Mol. Biol.* **288**, 765–786.

Miyazawa, A., Fujiyoshi, Y. & Unwin, N. (2003). *Structure and gating mechanism of the acetylcholine receptor pore*. *Nature (London)*, **423**, 949–955.

Morgan, D. G. & DeRosier, D. (1992). *Processing images of helical structures: a new twist*. *Ultramicroscopy*, **46**, 263–285.

Morris, D. M. & Jensen, G. J. (2008). *Toward a biomechanical understanding of whole bacterial cells*. *Annu. Rev. Biochem.* **77**, 583–613.

Mouche, F., Zhu, Y., Pulokas, J., Potter, C. S. & Carragher, B. (2003). *Automated three-dimensional reconstruction of keyhole limpet hemocyanin type 1*. *J. Struct. Biol.* **144**, 301–312.

Muench, S. P., Huss, M., Song, C. F., Phillips, C., Wieczorek, H., Trinick, J. & Harrison, M. A. (2009). *Cryo-electron microscopy of the vacuolar ATPase motor reveals its mechanical and regulatory complexity*. *J. Mol. Biol.* **386**, 989–999.

Muller, S. A., Aebi, U. & Engel, A. (2008). *What transmission electron microscopes can visualize now and in the future*. *J. Struct. Biol.* **163**, 235–245.

Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J. B., Engel, A. & Fujiyoshi, Y. (2000). *Structural determinants of water permeation through aquaporin-1*. *Nature (London)*, **407**, 599–605.

Nagayama, K. & Danev, R. (2008). *Phase contrast electron microscopy: development of thin-film phase plates and biological applications*. *Philos. Trans. R. Soc. London Ser. B*, **363**, 2153–2162.

Navaza, J. (2003). *On the three-dimensional reconstruction of icosahedral particles*. *J. Struct. Biol.* **144**, 13–23.

Navaza, J., Lepault, J., Rey, F. A., Álvarez-Rúa, C. & Borge, J. (2002). *On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation*. *Acta Cryst.* D**58**, 1820–1825.

Nicastro, D., McIntosh, J. R. & Baumeister, W. (2005). *3D structure of eukaryotic flagella in a quiescent state revealed by cryo-electron tomography*. *Proc. Natl Acad. Sci. USA*, **102**, 15889–15894.

Nicastro, D., Schwartz, C., Pierson, J., Gaudette, R., Porter, M. E. & McIntosh, J. R. (2006). *The molecular architecture of axonemes revealed by cryoelectron tomography*. *Science*, **313**, 944–948.

Nickell, S., Beck, F., Korinek, A., Mihalache, O., Baumeister, W. & Plitzko, J. M. (2007). *Automated cryoelectron microscopy of 'single particles' applied to the 26S proteasome*. *FEBS Lett.* **581**, 2751–2756.

Nickell, S., Forster, F., Linaroudis, A., Net, W. D., Beck, F., Hegerl, R., Baumeister, W. & Plitzko, J. M. (2005). *TOM software toolbox: acquisition and analysis for electron tomography*. *J. Struct. Biol.* **149**, 227–234.

Nogales, E., Wolf, S. G. & Downing, K. H. (1998). *Structure of the $\alpha\beta$ tubulin dimer by electron crystallography*. *Nature (London)*, **391**, 199–203.

Ogura, T. & Sato, C. (2006). *A fully automatic 3D reconstruction method using simulated annealing enables accurate posterioric angular assignment of protein projections*. *J. Struct. Biol.* **156**, 371–386.

Olins, D. E., Olins, A. L., Levy, H. A., Durfee, R. C., Margle, S. M., Tinnel, E. P. & Dover, S. D. (1983). *Electron microscopy tomography: transcription in three dimensions*. *Science*, **220**, 498–500.

Olson, N. H. & Baker, T. S. (1989). *Magnification calibration and the determination of spherical virus diameters using cryo-microscopy*. *Ultramicroscopy*, **30**, 281–298.

Orlova, A., Garner, E. C., Galkin, V. E., Heuser, J., Mullins, R. D. & Egelman, E. H. (2007). *The structure of bacterial ParM filaments*. *Nat. Struct. Mol. Biol.* **14**, 921–926.

Orzechowski, M. & Tama, F. (2008). *Flexible fitting of high-resolution X-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations*. *Biophys. J.* **95**, 5692–5705.

Ottmann, C., Marco, S., Jaspert, N., Marcon, C., Schauer, N., Weyand, M., Vandermeeren, C., Duby, G., Boutry, M., Wittinghofer, A., Rigaud, J. L. & Oecking, C. (2007). *Structure of a 14-3-3 coordinated hexamer of the plant plasma membrane H^+-ATPase by combining X-ray crystallography and electron cryomicroscopy*. *Mol. Cell*, **25**, 427–440.

Owen, C. H., Morgan, D. G. & DeRosier, D. J. (1996). *Image analysis of helical objects: the Brandeis helical package*. *J. Struct. Biol.* **116**, 167–175.

Penczek, P., Radermacher, M. & Frank, J. (1992). *Three-dimensional reconstruction of single particles embedded in ice*. *Ultramicroscopy*, **40**, 33–53.

Penczek, P. A., Grassucci, R. A. & Frank, J. (1994). *The ribosome at improved resolution: new techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles*. *Ultramicroscopy*, **53**, 251–270.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. (2004). *UCSF Chimera – a visualization system for exploratory research and analysis*. *J. Comput. Chem.* **25**, 1605–1612.

Plaisier, J. R., Jiang, L. & Abrahams, J. P. (2007). *Cyclops: new modular software suite for cryo-EM*. *J. Struct. Biol.* **157**, 19–27.

Polyakov, A., Richter, C., Malhotra, A., Koulich, D., Borukhov, S. & Darst, S. A. (1998). *Visualization of the binding site for the transcript cleavage factor GreB on Escherichia coli RNA polymerase*. *J. Mol. Biol.* **281**, 465–473.

Pomfret, A. J., Rice, W. J. & Stokes, D. L. (2007). *Application of the iterative helical real-space reconstruction method to large membranous tubular crystals of P-type ATPases*. *J. Struct. Biol.* **157**, 106–116.

Potter, C. S., Chu, H., Frey, B., Green, C., Kisseberth, N., Madden, T. J., Miller, K. L., Nahrstedt, K., Pulokas, J., Reilein, A., Tcheng, D., Weber, D. & Carragher, B. (1999). *Leginon: a system for fully automated acquisition of 1000 electron micrographs a day*. *Ultramicroscopy*, **77**, 153–161.

Prust, C. J., Doerschuk, P. C., Lander, G. C. & Johnson, J. E. (2009). *Ab initio maximum likelihood reconstruction from cryo electron microscopy images of an infectious virion of the tailed bacteriophage P22 and maximum likelihood versions of Fourier shell correlation appropriate for measuring resolution of spherical or cylindrical objects*. *J. Struct. Biol.* **167**, 185–199.

Rabl, J., Smith, D. M., Yu, Y., Chang, S. C., Goldberg, A. L. & Cheng, Y. (2008). *Mechanism of gate opening in the 20S proteasome by the proteasomal ATPases*. *Mol. Cell*, **30**, 360–368.

Radermacher, M. (1988). *Three-dimensional reconstruction of single particles from random and nonrandom tilt series*. *J. Electron Microsc. Tech.* **9**, 359–394.

Radermacher, M. (1991). *Three-dimensional reconstruction of single particles in electron microscopy*. In *Image Analysis in Biology*, edited by D.-P. Hader, pp. 219–246. Boca Raton: CRC Press.

Radermacher, M. (1992). *Weighted back-projection methods*. In *Electron Tomography*, edited by J. Frank, pp. 91–115. New York: Plenum Press.

Radermacher, M. (1994). *Three-dimensional reconstruction from random projections: orientational alignment via Radon transforms*. *Ultramicroscopy*, **53**, 121–136.

Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. (1987). *Three-dimensional reconstructions from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli*. *J. Microsc.* **146**, 113–136.

Ramey, V. H., Wang, H. W. & Nogales, E. (2009). *Ab initio reconstruction of helical samples with heterogeneity, disorder and coexisting symmetries*. *J. Struct. Biol.* **167**, 97–105.

Ranson, N. A., Clare, D. K., Farr, G. W., Houldershaw, D., Horwich, A. L. & Saibil, H. R. (2006). *Allosteric signaling of ATP hydrolysis in GroEL-GroES complexes*. *Nat. Struct. Mol. Biol.* **13**, 147–152.

Rayment, I., Holden, H. M., Whittaker, M., Yohn, C. B., Lorenz, M., Holmes, K. C. & Milligan, R. A. (1993). *Structure of the actin-myosin complex and its implications for muscle contraction*. *Science*, **261**, 58–65.

Reichow, S. L. & Gonen, T. (2009). *Lipid-protein interactions probed by electron crystallography*. *Curr. Opin. Struct. Biol.* **19**, 560–565.

Reimer, L. & Kohl, H. (2008). *Transmission Electron Microscopy: Physics of Image Formation*. New York: Springer Verlag.

Renault, L., Chou, H. T., Chiu, P. L., Hill, R. M., Zeng, X., Gipson, B., Zhang, Z. Y., Cheng, A., Unger, V. & Stahlberg, H. (2006). *Milestones in electron crystallography*. *J. Comput. Aided Mol. Des.* **20**, 519–527.

Roseman, A. M. (2000). *Docking structures of domains into maps from cryo-electron microscopy using local correlation*. *Acta Cryst.* D**56**, 1332–1340.

Roseman, A. M. (2003). *Particle finding in electron micrographs using a fast local correlation algorithm*. *Ultramicroscopy*, **94**, 225–236.

Rosenthal, P. B. & Henderson, R. (2003). *Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy*. *J. Mol. Biol.* **333**, 721–745.

Rossmann, M. G., Bernal, R. & Pletnev, S. V. (2001). *Combining electron microscopic with X-ray crystallographic structures*. *J. Struct. Biol.* **136**, 190–200.

Rubinstein, J. L., Walker, J. E. & Henderson, R. (2003). *Structure of the mitochondrial ATP synthase by electron cryomicroscopy*. *EMBO J.* **22**, 6182–6192.

Sachse, C., Chen, J. Z., Coureux, P. D., Stroupe, M. E., Fandrich, M. & Grigorieff, N. (2007). *High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus*. *J. Mol. Biol.* **371**, 812–835.

Salje, J., Zuber, B. & Lowe, J. (2009). *Electron cryomicroscopy of E. coli reveals filament bundles involved in plasmid DNA segregation*. *Science*, **323**, 509–512.

Sander, B., Golas, M. M. & Stark, H. (2005). *Advantages of CCD detectors for de novo three-dimensional structure determination in single-particle electron microscopy*. *J. Struct. Biol.* **151**, 92–105.

Sanner, M. F. (1999). *Python: A programming language for software integration and development*. *J. Mol. Graph. Model.* **17**, 57–61.

Sanner, M. F. (2005). *A component-based software environment for visualizing large macromolecular assemblies*. *Structure*, **13**, 447–462.

Schatz, M. & van Heel, M. (1990). *Invariant classification of molecular views in electron micrographs*. *Ultramicroscopy*, **32**, 255–264.

Schatz, M., Orlova, E. V., Dube, P., Jager, J. & van Heel, M. (1995). *Structure of Lumbricus terrestris hemoglobin at 30 Å resolution determined using angular reconstitution*. *J. Struct. Biol.* **114**, 28–40.

Scheres, S. H., Valle, M., Nunez, R., Sorzano, C. O., Marabini, R., Herman, G. T. & Carazo, J. M. (2005). *Maximum-likelihood multi-reference refinement for electron microscopy images*. *J. Mol. Biol.* **348**, 139–149.

Schmidt-Krey, I. (2007). *Electron crystallography of membrane proteins: two-dimensional crystallization and screening by electron microscopy*. *Methods*, **41**, 417–426.

Shacham, E., Sheehan, B. & Volkmann, N. (2007). *Density-based score for selecting near-native atomic models of unknown structures*. *J. Struct. Biol.* **158**, 188–195.

Shaikh, T. R., Barnard, D., Meng, X. & Wagenknecht, T. (2009). *Implementation of a flash-photolysis system for time-resolved cryo-electron microscopy*. *J. Struct. Biol.* **165**, 184–189.

Shaikh, T. R., Gao, H., Baxter, W. T., Asturias, F. J., Boisset, N., Leith, A. & Frank, J. (2008). *SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs*. *Nat. Protoc.* **3**, 1941–1974.

Sherman, M. B., Brink, J. & Chiu, W. (1996). *Performance of a slow-scan CCD camera for macromolecular imaging in a 400 kV electron cryomicroscope*. *Micron*, **27**, 129–139.

Shi, J., Williams, D. R. & Stewart, P. L. (2008). *A script-assisted microscopy (SAM) package to improve data acquisition rates on FEI Tecnai electron microscopes equipped with Gatan CCD cameras*. *J. Struct. Biol.* **164**, 166–169.

Short, J. M. (2004). *SLEUTH – a fast computer program for automatically detecting particles in electron microscope images*. *J. Struct. Biol.* **145**, 100–110.

Siegel, D. P. & Epand, R. M. (1997). *The mechanism of lamellar-to-inverted hexagonal phase transitions in phosphatidylethanolamine: implications for membrane fusion mechanisms*. *Biophys. J.* **73**, 3089–3111.

Siegel, D. P., Green, W. J. & Talmon, Y. (1994). *The mechanism of lamellar-to-inverted hexagonal phase transitions: a study using temperature-jump cryo-electron microscopy*. *Biophys. J.* **66**, 402–414.

Sorzano, C. O., Jonic, S., Nunez-Ramirez, R., Boisset, N. & Carazo, J. M. (2007). *Fast, robust, and accurate determination of transmission electron microscopy contrast transfer function*. *J. Struct. Biol.* **160**, 249–262.

Sorzano, C. O., Recarte, E., Alcorlo, M., Bilbao-Castro, J. R., San-Martin, C., Marabini, R. & Carazo, J. M. (2009). *Automatic particle selection from electron micrographs using machine learning techniques*. *J. Struct. Biol.* **167**, 252–260.

Sosa, H., Hoenger, A. & Milligan, R. A. (1997). *Three different approaches for calculating the three-dimensional structure of microtubules decorated with kinesin motor domains*. *J. Struct. Biol.* **118**, 149–158.

Sosa, H. & Milligan, R. A. (1996). *Three-dimensional structure of ncd-decorated microtubules obtained by a back-projection method*. *J. Mol. Biol.* **260**, 743–755.

Spence, J. C. H. (2009). *High-Resolution Electron Microscopy*. Oxford University Press.

Stagg, S. M., Lander, G. C., Pulokas, J., Fellmann, D., Cheng, A., Quispe, J. D., Mallick, S. P., Avila, R. M., Carragher, B. & Potter, C. S. (2006). *Automated cryoEM data acquisition and analysis of 284742 particles of GroEL*. *J. Struct. Biol.* **155**, 470–481.

Stagg, S. M., LaPointe, P., Razvi, A., Gurkan, C., Potter, C. S., Carragher, B. & Balch, W. E. (2008). *Structural basis for cargo regulation of COPII coat assembly*. *Cell*, **134**, 474–484.

Steven, A. C. & Baumeister, W. (2008). *The future is hybrid*. *J. Struct. Biol.* **163**, 186–195.

Stewart, M. (1988). *Computer image processing of electron micrographs of biological structures with helical symmetry*. *J. Electron Microsc. Tech.* **9**, 325–358.

Stewart, P. L., Chiu, C. Y., Huang, S., Muir, T., Zhao, Y., Chait, B., Mathias, P. & Nemerow, G. R. (1997). *Cryo-EM visualization of an exposed RGD epitope on adenovirus that escapes antibody neutralization*. *EMBO J.* **16**, 1189–1198.

Stewart, P. L., Fuller, S. D. & Burnett, R. M. (1993). *Difference imaging of adenovirus: bridging the resolution gap between X-ray crystallography and electron microscopy*. *EMBO J.* **12**, 2589–2599.

Subramaniam, S., Gerstein, M., Oesterhelt, D. & Henderson, R. (1993). *Electron diffraction analysis of structural changes in the photocycle of bacteriorhodopsin*. *EMBO J.* **12**, 1–18.

Suhre, K., Navaza, J. & Sanejouand, Y.-H. (2006). *NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps*. *Acta Cryst.* D**62**, 1098–1100.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C. S. & Carragher, B. (2005). *Automated molecular microscopy: the new Leginon system*. *J. Struct. Biol.* **151**, 41–60.

Suloway, C., Shi, J., Cheng, A., Pulokas, J., Carragher, B., Potter, C. S., Zheng, S. Q., Agard, D. A. & Jensen, G. J. (2009). *Fully automated, sequential tilt-series acquisition with Leginon*. *J. Struct. Biol.* **167**, 11–18.

Tama, F., Miyashita, O. & Brooks, C. L. III (2004*a*). *Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis*. *J. Mol. Biol.* **337**, 985–999.

Tama, F., Miyashita, O. & Brooks, C. L. III (2004*b*). *Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM*. *J. Struct. Biol.* **147**, 315–326.

Tama, F., Valle, M., Frank, J. & Brooks, C. L. III (2003). *Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy*. *Proc. Natl Acad. Sci. USA*, **100**, 9319–9323.

Tan, R. K., Devkota, B. & Harvey, S. C. (2008). *YUP.SCX: coaxing atomic models into medium resolution electron density maps*. *J. Struct. Biol.* **163**, 163–174.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. (2007). *EMAN2: an extensible image processing suite for electron microscopy*. *J. Struct. Biol.* **157**, 38–46.

Tang, J., Olson, N., Jardine, P. J., Grimes, S., Anderson, D. L. & Baker, T. S. (2008). *DNA poised for release in bacteriophage φ29*. *Structure*, **16**, 935–943.

Tao, Y., Olson, N. H., Xu, W., Anderson, D. L., Rossmann, M. G. & Baker, T. S. (1998). *Assembly of a tailed bacterial virus and its genome release studied in three dimensions*. *Cell*, **95**, 431–437.

Taylor, K. A. & Glaeser, R. M. (1974). *Electron diffraction of frozen, hydrated protein crystals*. *Science*, **186**, 1036–1037.

Taylor, K. A., Tang, J., Cheng, Y. & Winkler, H. (1997). *The use of electron tomography for structural analysis of disordered protein arrays*. *J. Struct. Biol.* **120**, 372–386.

Thuman-Commike, P. A. & Chiu, W. (1997). *Improved common line-based icosahedral particle image orientation estimation algorithms*. *Ultramicroscopy*, **68**, 231–255.

Thuman-Commike, P. A & Chiu, W. (2000). *Reconstruction principles of icosahedral virus structure determination using electron cryomicroscopy*. *Micron*, **31**, 687–711.

Tichelaar, W., Oostergetel, G. T., Haker, J., van Heel, M. G. & van Bruggen, E. F. J. (1980). *Scanning transmission electron microscopy of biological macromolecules*. *Ultramicroscopy*, **5**, 27–33.

Topf, M., Baker, M. L., John, B., Chiu, W. & Sali, A. (2005). *Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy*. *J. Struct. Biol.* **149**, 191–203.

Topf, M., Baker, M. L., Marti-Renom, M. A., Chiu, W. & Sali, A. (2006). *Refinement of protein structures by iterative comparative modeling and cryoEM density fitting*. *J. Mol. Biol.* **357**, 1655–1668.

Toyoshima, C. & Unwin, N. (1990). *Three-dimensional structure of the acetylcholine receptor by cryoelectron microscopy and helical image reconstruction*. *J. Cell Biol.* **111**, 2623–2635.

Toyoshima, C., Yonekura, K. & Sasabe, H. (1993). *Contrast transfer for frozen-hydrated specimens. II. Amplitude contrast at very low frequencies*. *Ultramicroscopy*, **48**, 165–176.

Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. (2008). *Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics*. *Structure*, **16**, 673–683.

Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. (2009). *Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography*. *Methods*, **49**, 174–180.

Trachtenberg, S. (1998). *A fast-freezing device with a retractable environmental chamber, suitable for kinetic cryo-electron microscopy studies*. *J. Struct. Biol.* **123**, 45–55.

Turchetta, R., Allport, P. P., Casse, G., Clark, A., Crooks, J., Evans, A., Fant, A., Faruqi, A. R., French, M. J. & Henderson, R. (2006). *CMOS monolithic active pixel sensors (MAPS): new 'eyes' for science*. *Nucl. Instrum. Methods Phys. Res. A*, **560**, 139–142.

Unger, V. M. (2000). *Assessment of electron crystallographic data obtained from two-dimensional crystals of biological specimens*. *Acta Cryst.* D**56**, 1259–1269.

Unser, M., Trus, B. L., Frank, J. & Steven, A. C. (1989). *The spectral signal-to-noise ratio resolution criterion: computational efficiency and statistical precision*. *Ultramicroscopy*, **30**, 429–434.

Unwin, N. (1993). *Nicotinic acetylcholine receptor at 9 Å resolution*. *J. Mol. Biol.* **229**, 1101–1124.

Unwin, N. (1995). *Acetylcholine receptor channel imaged in the open state*. *Nature (London)*, **373**, 37–43.

Unwin, P. N. T. & Henderson, R. (1975). *Molecular structure determination by electron microscopy of unstained crystalline specimens*. *J. Mol. Biol.* **94**, 425–440.

Valpuesta, J. M., Carrascosa, J. L. & Henderson, R. (1994). *Analysis of electron microscope images and electron diffraction patterns of thin crystals of Φ29 connectors in ice*. *J. Mol. Biol.* **240**, 281–287.

Velazquez-Muriel, J. A. & Carazo, J. M. (2007). *Flexible fitting in 3D-EM with incomplete data on superfamily variability*. *J. Struct. Biol.* **158**, 165–181.

Velazquez-Muriel, J. A., Valle, M., Santamaria-Pang, A., Kakadiaris, I. A. & Carazo, J. M. (2006). *Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies*. *Structure*, **14**, 1115–1126.

Venien-Bryan, C., Jonic, S., Skamnaki, V., Brown, N., Bischler, N., Oikonomakos, N. G., Boisset, N. & Johnson, L. N. (2009). *The structure of phosphorylase kinase holoenzyme at 9.9 Å resolution and location of the catalytic subunit and the substrate glycogen phosphorylase*. *Structure*, **17**, 117–127.

Villa, E., Sengupta, J., Trabuco, L. G., LeBarron, J., Baxter, W. T., Shaikh, T. R., Grassucci, R. A., Nissen, P., Ehrenberg, M., Schulten, K. & Frank, J. (2009). *Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis*. *Proc. Natl Acad. Sci. USA*, **106**, 1063–1068.

Volkmann, N. (2002). *A novel three-dimensional variant of the watershed transform for segmentation of electron density maps*. *J. Struct. Biol.* **138**, 123–129.

Volkmann, N. & Hanein, D. (1999). *Quantitative fitting of atomic models into observed densities derived by electron microscopy*. *J. Struct. Biol.* **125**, 176–184.

Volkmann, N. & Hanein, D. (2003). *Docking of atomic models into reconstructions from electron microscopy*. *Methods Enzymol.* **374**, 204–225.

Voss, N. R., Yoshioka, C. K., Radermacher, M., Potter, C. S. & Carragher, B. (2009). *DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy*. *J. Struct. Biol.* **166**, 205–213.

Wade, R. H. (1992). *A brief look at imaging and contrast transfer*. *Ultramicroscopy*, **46**, 145–156.

Wade, R. H. & Frank, J. (1977). *Electron microscope transfer functions for partially coherent axial illumination and chromatic defocus spread*. *Optik*, **49**, 81–92.

Walker, M., Zhang, X.-Z., Jiang, W., Trinick, J. & White, H. D. (1999). *Observation of transient disorder during myosin subfragment-1 binding to actin by stopped-flow fluorescence and millisecond time resolution electron cryomicroscopy: evidence that the start of the crossbridge power stroke in muscle has variable geometry*. *Proc. Natl Acad. Sci. USA*, **96**, 465–470.

Walz, J., Tamura, T., Tamura, N., Grimm, R., Baumeister, W. & Koster, A. J. (1997). *Tricorn protease exists as an icosahedral supermolecule in vivo*. *Mol. Cell*, **1**, 59–65.

Wang, G. J., Porta, C., Chen, Z., Baker, T. S. & Johnson, J. E. (1992). *Identification of a Fab interaction footprint site on an icosahedral virus by cryoelectron microscopy and X-ray crystallography*. *Nature (London)*, **355**, 275–278.

Wang, H. W. & Nogales, E. (2005). *An iterative Fourier–Bessel algorithm for reconstruction of helical structures with severe Bessel overlap*. *J. Struct. Biol.* **149**, 65–78.

Wang, M. C., Collins, R. F., Ford, R. C., Berrow, N. S., Dolphin, A. C. & Kitmitto, A. (2004). *The three-dimensional structure of the cardiac L-type voltage-gated calcium channel: comparison with the skeletal muscle form reveals a common architectural motif*. *J. Biol. Chem.* **279**, 7159–7168.

Wang, Y. A., Yu, X., Overman, S., Tsuboi, M., Thomas, G. J. Jr & Egelman, E. H. (2006). *The structure of a filamentous bacteriophage*. *J. Mol. Biol.* **361**, 209–215.

Weik, M., Ravelli, R. B., Kryger, G., McSweeney, S., Raves, M. L., Harel, M., Gros, P., Silman, I., Kroon, J. & Sussman, J. L. (2000). *Specific chemical and structural damage to proteins produced by synchrotron radiation*. *Proc. Natl Acad. Sci. USA*, **97**, 623–628.

White, H. D., Thirumurugan, K., Walker, M. L. & Trinick, J. (2003). *A second generation apparatus for time-resolved electron cryomicroscopy using stepper motors and electrospray*. *J. Struct. Biol.* **144**, 246–252.

White, H. D., Walker, M. L. & Trinick, J. (1998). *A computer-controlled spraying-freezing apparatus for millisecond time-resolution electron cryomicroscopy*. *J. Struct. Biol.* **121**, 306–313.

Whittaker, M., Carragher, B. O. & Milligan, R. A. (1995). *PHOELIX: a package for semi-automated helical reconstruction*. *Ultramicroscopy*, **58**, 245–259.

Wilson-Kubalek, E. M., Brown, R. E., Celia, H. & Milligan, R. A. (1998). *Lipid nanotubes as substrates for helical crystallization of macromolecules*. *Proc. Natl Acad. Sci. USA*, **95**, 8040–8045.

Winkelmann, D. A., Baker, T. S. & Rayment, I. (1991). *Three-dimensional structure of myosin subfragment-1 from electron microscopy of sectioned crystals*. *J. Cell Biol.* **114**, 701–713.

Winkler, H. & Taylor, K. A. (1994). *3D reconstruction by combining data from sections cut oblique to different unit cell axes*. *Ultramicroscopy*, **55**, 357–371.

Winkler, H. & Taylor, K. A. (1996). *Software for 3-D reconstruction from images of oblique sections through 3-D crystals*. *J. Struct. Biol.* **116**, 241–247.

Woolford, D., Ericksson, G., Rothnagel, R., Muller, D., Landsberg, M. J., Pantelic, R. S., McDowall, A., Pailthorpe, B., Young, P. R., Hankamer, B. & Banks, J. (2007). *SwarmPS: rapid, semi-automated single particle selection software*. *J. Struct. Biol.* **157**, 174–188.

Wriggers, W. & Birmanns, S. (2001). *Using Situs for flexible and rigid-body fitting of multiresolution single-molecule data*. *J. Struct. Biol.* **133**, 193–202.

Wriggers, W. & Chacon, P. (2001). *Modeling tricks and fitting techniques for multiresolution structures*. *Structure*, **9**, 779–788.

Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). *Situs: a package for docking crystal structures into low-resolution maps from electron microscopy*. *J. Struct. Biol.* **125**, 185–195.

Xuong, N. H., Jin, L., Kleinfelder, S., Li, S., Leblanc, P., Duttweiler, F., Bouwer, J. C., Peltier, S. T., Milazzo, A. C. & Ellisman, M. (2007). *Future directions for camera systems in electron microscopy*. *Methods Cell Biol.* **79**, 721–739.

Yan, X., Dryden, K. A., Tang, J. & Baker, T. S. (2007). *Ab initio random model method facilitates 3D reconstruction of icosahedral particles*. *J. Struct. Biol.* **157**, 211–225.

Yan, X., Sinkovits, R. S. & Baker, T. S. (2007). *AUTO3DEM – an automated and high throughput program for image reconstruction of icosahedral particles*. *J. Struct. Biol.* **157**, 73–82.

Yang, C., Penczek, P. A., Leith, A., Asturias, F. J., Ng, E. G., Glaeser, R. M. & Frank, J. (2007). *The parallelization of SPIDER on distributed-memory computers using MPI*. *J. Struct. Biol.* **157**, 240–249.

Yeager, M., Unger, V. M. & Mitra, A. K. (1999). *Three-dimensional structure of membrane proteins determined by two-dimensional crystallization, electron cryomicroscopy, and image analysis*. *Methods Enzymol.* **294**, 135–180.

Yonekura, K., Braunfeld, M. B., Maki-Yonekura, S. & Agard, D. A. (2006). *Electron energy filtering significantly improves amplitude contrast of frozen-hydrated protein at 300kV*. *J. Struct. Biol.* **156**, 524–536.

Yonekura, K., Maki-Yonekura, S. & Namba, K. (2003). *Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy*. *Nature (London)*, **424**, 643–650.

Yonekura, K., Toyoshima, C., Maki-Yonekura, S. & Namba, K. (2003). *GUI programs for processing individual images in early stages of helical image reconstruction – for high-resolution structure analysis*. *J. Struct. Biol.* **144**, 184–194.

Yoshimura, H., Matsumoto, M., Endo, S. & Nagayama, K. (1990). *Two-dimensional crystallization of proteins on mercury*. *Ultramicroscopy*, **32**, 265–274.

Yoshioka, C., Pulokas, J., Fellmann, D., Potter, C. S., Milligan, R. A. & Carragher, B. (2007). *Automation of random conical tilt and orthogonal tilt data collection using feature-based correlation*. *J. Struct. Biol.* **159**, 335–346.

Yu, X., Hiromasa, Y., Tsen, H., Stoops, J. K., Roche, T. E. & Zhou, Z. H. (2008). *Structures of the human pyruvate dehydrogenase complex cores: a highly conserved catalytic center with flexible N-terminal domains*. *Structure*, **16**, 104–114.

Yu, X., Jin, L. & Zhou, Z. H. (2008). *3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy*. *Nature (London)*, **453**, 415–419.

Yu, X., Wang, L., Acehan, D., Wang, X. & Akey, C. W. (2006). *Three-dimensional structure of a double apoptosome formed by the Drosophila Apaf-1 related killer*. *J. Mol. Biol.* **355**, 577–589.

Zanetti, G., Briggs, J. A., Grunewald, K., Sattentau, Q. J. & Fuller, S. D. (2006). *Cryo-electron tomographic structure of an immunodeficiency virus envelope complex in situ*. *PLoS Pathog.* **2**, 790–797.

Zanetti, G., Riches, J. D., Fuller, S. D. & Briggs, J. A. (2009). *Contrast transfer function correction applied to cryo-electron tomography and sub-tomogram averaging*. *J. Struct. Biol.* **168**, 305–312.

Zemlin, F. (1992). *Desired features of a cryoelectron microscope for the electron crystallography of biological material*. *Ultramicroscopy*, **46**, 25–32.

Zemlin, F., Beckmann, E. & van der Mast, K. D. (1996). *A 200 kV electron microscope with Schottky field emitter and a helium-cooled superconducting objective lens*. *Ultramicroscopy*, **63**, 227–238.

Zeng, X., Gipson, B., Zheng, Z. Y., Renault, L. & Stahlberg, H. (2007). *Automatic lattice determination for two-dimensional crystal images*. *J. Struct. Biol.* **160**, 353–361.

Zhang, J., Nakamura, N., Shimizu, Y., Liang, N., Liu, X., Jakana, J., Marsh, M. P., Booth, C. R., Shinkawa, T., Nakata, M. & Chiu, W. (2009). *JADAS: a customizable automated data acquisition system and its application to ice-embedded single particles*. *J. Struct. Biol.* **165**, 1–9.

Zhang, P., Borgnia, M. J., Mooney, P., Shi, D., Pan, M., O'Herron, P., Mao, A., Brogan, D., Milne, J. L. & Subramaniam, S. (2003). *Automated image acquisition and processing using a new generation of 4K × 4K CCD cameras for cryo electron microscopic studies of macromolecular assemblies*. *J. Struct. Biol.* **143**, 135–144.

Zhang, X., Settembre, E., Xu, C., Dormitzer, P. R., Bellamy, R., Harrison, S. C. & Grigorieff, N. (2008). *Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction*. *Proc. Natl Acad. Sci. USA*, **105**, 1867–1872.

Zheng, Q. S., Braunfeld, M. B., Sedat, J. W. & Agard, D. A. (2004). *An improved strategy for automated electron microscopic tomography*. *J. Struct. Biol.* **147**, 91–101.

Zheng, S. Q., Keszthelyi, B., Branlund, E., Lyle, J. M., Braunfeld, M. B., Sedat, J. W. & Agard, D. A. (2007). *UCSF tomography: an integrated software suite for real-time electron microscopic tomographic data collection, alignment, and reconstruction*. *J. Struct. Biol.* **157**, 138–147.

Zheng, S. Q., Kollman, J. M., Braunfeld, M. B., Sedat, J. W. & Agard, D. A. (2006). *Automated acquisition of electron microscopic random conical tilt sets*. *J. Struct. Biol.* **157**, 148–155.

Zhou, Z. H. (2008). *Towards atomic resolution structural determination by single-particle cryo-electron microscopy*. *Curr. Opin. Struct. Biol.* **18**, 218–228.

Zhou, Z. H., Baker, M. L., Jiang, W., Dougherty, M., Jakana, J., Dong, G., Lu, G. & Chiu, W. (2001). *Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus*. *Nat. Struct. Biol.* **8**, 868–873.

Zhou, Z. H. & Chiu, W. (1993). *Prospects for using an IVEM with a FEG for imaging macromolecules towards atomic resolution*. *Ultramicroscopy*, **49**, 407–416.

Zhou, Z. H., Chiu, W., Haskell, K., Spears, H. J., Jakana, J., Rixon, F. J. & Scott, L. R. (1998). *Refinement of herpesvirus B-capsid structure on parallel supercomputers*. *Biophys. J.* **74**, 576–588.

Zhou, Z. H., Hardt, S., Wang, B., Sherman, M. B., Jakana, J. & Chiu, W. (1996). *CTF determination of images of ice-embedded single particles using a graphics interface*. *J. Struct. Biol.* **116**, 216–222.

Zhu, P., Winkler, H., Chertova, E., Taylor, K. A. & Roux, K. H. (2008). *Cryoelectron tomography of HIV-1 envelope spikes: further evidence for tripod-like legs*. *PLoS Pathog.* **4**, e1000203.

Zhu, Y., Carragher, B., Glaeser, R. M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., de Haas, F., Hall, R. J., Kriegman, D. J., Ludtke, S. J., Mallick, S. P., Penczek, P. A., Roseman, A. M., Sigworth, F. J., Volkmann, N. & Potter, C. S. (2004). *Automatic particle selection: results of a comparative study*. *J. Struct. Biol.* **145**, 3–14.

Ziese, U., Geerts, W. J., van der Krift, T. P., Verkleij, A. J. & Koster, A. J. (2003). *Correction of autofocusing errors due to specimen tilt for automated electron tomography*. *J. Microsc.* **211**, 179–185.

Ziese, U., Janssen, A. H., Murk, J. L., Geerts, W. J., van der Krift, T., Verkleij, A. J. & Koster, A. J. (2002). *Automated high-throughput electron tomography by pre-calibration of image shifts*. *J. Microsc.* **205**, 187–200.

Zuber, B., Chami, M., Houssin, C., Dubochet, J., Griffiths, G. & Daffe, M. (2008). *Direct visualization of the outer membrane of mycobacteria and corynebacteria in their native state*. *J. Bacteriol.* **190**, 5672–5680.

Zuber, B., Nikonenko, I., Klauser, P., Muller, D. & Dubochet, J. (2005). *The mammalian central nervous synaptic cleft contains a high density of periodically organized complexes*. *Proc. Natl Acad. Sci. USA*, **102**, 19192–19197.

International Tables for Crystallography (2012). Vol. F, Chapter 19.7, pp. 615–619.

Chapter 19.7. Nuclear magnetic resonance (NMR) spectroscopy

K. Wüthrich

19.7.1. Complementary roles of NMR in solution and X-ray crystallography in structural biology

X-ray diffraction in crystals and NMR in solution can both be used to determine the complete three-dimensional structure of biological macromolecules, and to date a significant number of globular protein structures have been independently determined in crystals and in solution (Billeter, 1992). Particularly detailed comparisons of the two states have been made for the α-amylase inhibitor Tendamistat, which also included solving the crystal structure by molecular replacement with the NMR structure (Braun *et al.*, 1989). The dominant impression is one of near-identity of the molecular architecture in solution and in single crystals, which holds for the polypeptide backbone as well as the core side chains.

Although the presently available results show that there is usually close coincidence between both the global molecular architecture and the detailed arrangement of the molecular core in corresponding X-ray and NMR structures of globular proteins, there is also extensive complementarity in the information that is accessible with the two methods: X-ray diffraction can provide the desired information for big molecules and multimolecular assemblies, whereas NMR structure determination is limited to smaller systems [recently introduced new experiments enable solution NMR measurements for molecular weights of 100 000 and beyond (Pervushin *et al.*, 1997; Riek *et al.*, 1999)]. NMR measurements in turn provide quantitative information on both very rapid motions on the subnanosecond timescale (Otting *et al.*, 1991; Peng & Wagner, 1992) and slower dynamic processes (Wüthrich, 1986) which are not manifested in the X-ray data. Examples of low-frequency intramolecular mobility are the ring flips of phenylalanine and tyrosine (Wüthrich & Wagner, 1975), exchange of interior hydration water molecules with the bulk solvent (Otting *et al.*, 1991), and interconversion of disulfide bonds between the *R* and *S* chiral forms (Otting *et al.*, 1993). NMR studies of amide proton exchange and *cis–trans* isomerization of *Xxx*—Pro peptide bonds (Wüthrich, 1976, 1986) afford additional insight into conformational equilibria in the protein core. Finally, in all instances where a biological macromolecule cannot be crystallized, NMR is currently the only method capable of providing a three-dimensional structure.

Overall, X-ray crystal structures and NMR solution structures provide qualitatively different information on the molecular surface. In the crystals, a sizeable proportion of surface amino-acid side chains are subject to similar packing constraints in protein–protein interfaces as the interior side chains in the protein core, and therefore they are rather precisely defined by the X-ray diffraction data. In NMR solution structures determined according to a standard protocol (Wüthrich, 1995), the surface is usually largely disordered. Surface disorder in NMR structures may in part arise from scarcity of nuclear Overhauser effect (NOE) distance constraints and packing constraints near the protein surface, but, in turn, scarcity of NOE constraints is often a direct consequence of dynamic disorder. Additional NMR experiments that are not part of a standard structure determination protocol can provide information needed for more detailed characterization of the molecular surface, but care must be exercised in the data analysis because of the presence of a multitude of equilibria between two or multiple transient local conformational states, of which the relative populations are usually not independently known.

19.7.2. A standard protocol for NMR structure determination of proteins and nucleic acids

An NMR structure determination involves sample preparation, NMR measurements, assignment of the NMR lines to individual atoms in the polymer chain, collection of conformational constraints, and structure calculation and refinement, where in present practice the sequence of steps usually corresponds to the flow diagram of Fig. 19.7.2.1. As is also indicated in Fig. 19.7.2.1, it is a special feature of protein structure determination by NMR that the secondary polypeptide structure, including the connections between individual segments of regular secondary structure, may be known early on from the data used for obtaining the resonance assignments, *i.e.* before the structure calculation is even started.

For the sample preparation, homogeneous macromolecular material is dissolved at about 1 m*M* concentration in 0.5 ml of water. The ionic strength, pH and temperature, and possibly the concentration of additives, may then be adjusted, for example, to ensure near-physiological conditions, or denaturing conditions *etc*. The NMR study will often include the preparation of compounds enriched with ^{15}N and/or ^{13}C, and possibly with ^{2}H (Kay & Gardner, 1997). Uniformly isotope-labelled recombinant proteins are routinely obtained by expression in *Escherichia coli* bacteria grown on minimal media. For RNA and DNA, isotope-labelling techniques are more involved, but labelled nucleic acids will also be commonly available in the future. Being able to work

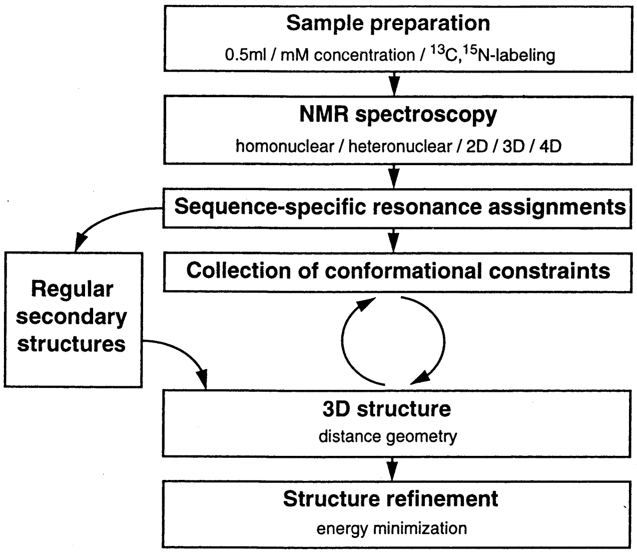

Sample preparation
0.5ml / mM concentration / ^{13}C,^{15}N-labeling
NMR spectroscopy
homonuclear / heteronuclear / 2D / 3D / 4D
Sequence-specific resonance assignments
Regular secondary structures
Collection of conformational constraints
3D structure
distance geometry
Structure refinement
energy minimization

Figure 19.7.2.1
Diagram outlining the course of a macromolecular structure determination by NMR in solution.

Copyright © 2012 International Union of Crystallography

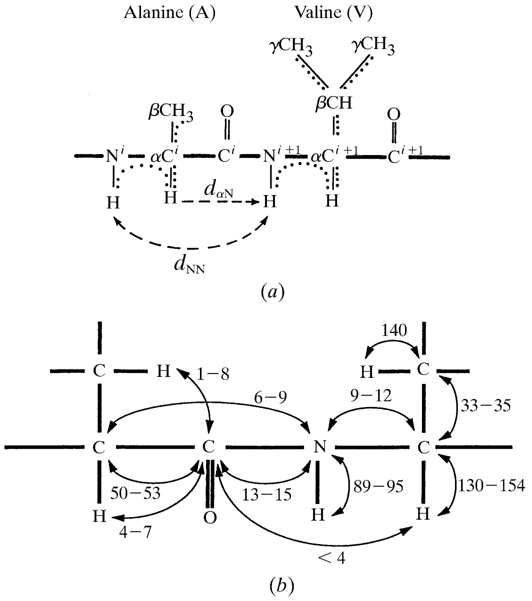


Figure 19.7.2.2

(*a*) Sequential resonance assignment based on sequential ^{1}H–^{1}H NOEs. In the dipeptide segment -Ala-Val- the dotted lines indicate ^{1}H–^{1}H relations which can be established by scalar through-bond spin–spin couplings. The broken arrows connect pairs of protons in sequentially neighbouring residues, *i* and $i + 1$, which are related by ^{1}H–^{1}H NOEs that manifest short sequential distances $d_{\alpha N}$ (between αCH and the amide proton of the following residue) and d_{NN} (between the amide protons of neighbouring residues). (*b*) Segment of a polypeptide chain with indication of the scalar spin–spin couplings that provide the basis for obtaining sequential assignments by triple-resonance experiments with uniformly ^{13}C/^{15}N labelled proteins.

with solutions is generally considered to be a great asset of the NMR method, but there are also potential inherent difficulties. For example, in the course of an investigation it may be nontrivial to achieve identical solution conditions in different NMR samples of the same compound, the absence of which typically results in small chemical-shift differences that slow down the combined analysis of different NMR spectra.

The demands on NMR experiments for macromolecular structure determination are currently met by multidimensional NMR at high polarizing magnetic fields (Wüthrich, 1986; Ernst *et al.*, 1987; Cavanagh *et al.*, 1996). With increasing molecular size and concomitant increase of the number of NMR peaks, it becomes more and more difficult to resolve and assign the individual resonances. In heteronuclear three- or four-dimensional (3D or 4D) spectra recorded with compounds that are uniformly labelled with ^{15}N and/or ^{13}C, the peaks are spread out in a third and possibly fourth dimension along the ^{15}N and ^{13}C chemical-shift axes (for a review, see Bax & Grzesiek, 1993). Additional labelling with ^{2}H results in slower spin-relaxation rates and hence improved sensitivity and spectral resolution. Although this approach provides ^{13}C and ^{15}N NMR information, which may be used to support the structure determination (*e.g.* Wishart *et al.*, 1991) and to provide supplementary information on dynamic features of the molecule studied, the key purpose of heteronuclear NMR experiments is to enhance the spectral resolution for studies of protons and thus to obtain the maximum possible number of ^{1}H NMR-based conformational constraints.

Resonance assignments in biopolymers are obtained using the sequential assignment strategy (Dubs *et al.*, 1979; Billeter *et al.*, 1982; Wagner & Wüthrich, 1982). Unique segments of two or several sequentially adjoining amino-acid residues are identified by NMR experiments and then attributed to discrete positions in the polypeptide chain by comparison with the chemically determined amino-acid sequence. The desired relations between protons in sequentially neighbouring amino-acid residues *i* and $i + 1$ can be established by nuclear Overhauser effects (NOEs), manifesting a close approach between αCH_i and NH_{i+1} ($d_{\alpha N}$), NH_i and NH_{i+1} (d_{NN}) (Fig. 19.7.2.2*a*), and possibly βCH_i and NH_{i+1} ($d_{\beta N}$) (Wüthrich, 1986). For *Xxx*—Pro dipeptide segments, corresponding connectivities are observed with δCH_2 of Pro in the place of the amide proton. For small proteins with up to about 100 amino-acid residues, NOE-based sequential assignments can rely entirely on homonuclear ^{1}H NMR experiments, and for somewhat bigger proteins they can be established based on the improved resolution of 3D NMR experiments with ^{15}N-labelled proteins. No prior knowledge of the polypeptide conformation is needed, since at least one of the two distances $d_{\alpha N}$ or d_{NN} (Fig. 19.7.2.2*a*) is always sufficiently short to be observed by NOEs (Billeter *et al.*, 1982). A further attractive feature of this approach is that the identification of the sequential NOEs forms an integral part of the data collection for the protein structure determination (see below). Sequential assignments can alternatively be obtained entirely *via* heteronuclear scalar couplings, using recombinant isotope-labelled proteins (Fig. 19.7.2.2*b*). Using 3D and 4D heteronuclear triple-resonance experiments, the resonance lines of sufficiently large mutually overlapping fragments of the polypeptide chain are grouped together to enable sequence-specific resonance assignments (for a review, see Bax & Grzesiek, 1993). With the implementation of transverse relaxation-optimized spectroscopy (TROSY) elements (Pervushin *et al.*, 1997) into triple-resonance experiments (Salzmann *et al.*, 1998), backbone resonance assignments *via* the spin–spin couplings of Fig. 19.7.2.2(*b*) can be performed with molecular weights of 100 000 and beyond. For nucleic acids, assignment procedures were largely patterned after those used for proteins and have been used successfully for fragments with 40 nucleotides and beyond.

NOE upper-distance constraints contain the crucial information needed for macromolecular structure determination (Wüthrich, 1986, 1989). To obtain a high-quality structure, the maximum possible number of NOE conformational constraints must be collected as input for the structure calculation. This is accomplished by using the chemical-shift lists obtained as a result of the sequence-specific resonance assignments to attribute the cross peaks in 2D [^{1}H, ^{1}H]-NOESY spectra, or 3D and 4D heteronuclear-resolved [^{1}H, ^{1}H]-NOESY spectra, to distinct pairs of hydrogen atoms. As indicated in Fig. 19.7.2.1, this data collection is achieved in several cycles, where ambiguities in the NOESY cross-peak assignments can usually be resolved by reference to preliminary structures calculated from incomplete input data sets (Güntert *et al.*, 1993). In present practice, each individual NOE constraint has the format of an allowed distance range, which circumvents intrinsic difficulties that might arise from attempts at quantitative distance measurements, and which is also adjusted to account for possible effects from internal mobility. The lower limit is usually taken to correspond to the sum of two hydrogen atomic radii, *i.e.* 2.0 Å, and the NOE intensities are translated into corresponding upper bounds, typically in steps of 2.5, 3.0 and 4.0 Å. Supplementary conformational constraints, for example, from spin–spin coupling constants (Wüthrich, 1986), residual dipole–dipole couplings (Tjandra & Bax, 1997), pseudocontact shifts and relaxation effects near paramagnetic centres (Banci *et al.*, 1998) *etc.*, are represented in the input by similar allowed ranges, which account for the internal mobility of the 3D structures and the limited accuracy of the individual measurements. Initially, NMR structures were calculated using distance-geometry techniques, and subsequently the principles of distance geometry have been introduced into molecular-dynamics programs in Cartesian coordinates (Brünger *et al.*, 1986) or in torsion-angle space (Güntert *et al.*, 1997). Model calculations performed in conjunction with the initial protein structure determinations had shown that NMR structure calculation depends critically on the density of NOE distance constraints, while it is remarkably robust with regard to low precision of the individual distance constraints (Havel & Wüthrich, 1985). For the common presentation of an NMR structure, one considers the result of a single structure calculation as representing one molecular geometry that is compatible with the NMR data. To investigate further whether or not this solution is unique, the calculation is repeated with different boundary conditions, where for each calculation, convergence is judged by the residual constraint violations. All satisfactory solutions, by this criterion, are included in a group of conformers that is used to represent the NMR structure (Fig. 19.7.2.3). The precision of the structure determination is reflected by the dispersion among this group of conformers. In proteins, larger variations are typically observed near the chain ends, in exposed loops and for surface amino-acid side chains, which contrasts with the well defined core. For nucleic acids, the 'global folds', for example, formation of duplexes, triplexes, quadruplexes, or loops, can be well defined by NMR, but because of the short range of the NOE distance measurements, certain 'long-range' features, for example, bending of DNA duplexes, may be more difficult to characterize.

19.7.3. Combined use of single-crystal X-ray diffraction and solution NMR for structure determination

The chemical shifts in proteins or nucleic acids cannot be calculated with sufficiently high precision from the X-ray crystal structures to predict the NMR spectra reliably. This limits the extent to which improved efficiency of an NMR structure determination may be derived from knowledge of the corresponding crystal structure. Nonetheless, for resonance assignments with sequential NOEs, information on the type of connectivity to look for in specified polypeptide segments may be derived. The X-ray structure may also tentatively be used as a starting reference to resolve ambiguities in the assignment of NOE distance constraints (Güntert *et al.*, 1993), which may help to reduce the number of cycles needed to collect the input data for a high-quality structure determination (Fig. 19.7.2.1).

An attractive possibility for combined use of X-ray and NMR data has in the past arisen in situations where both a low-resolution crystallographic electron-density map and a secondary-structure determination by NMR are available. As mentioned before, NMR secondary-structure determination in proteins usually derives directly from the backbone resonance assignments (Wishart *et al.*, 1991; Wüthrich *et al.*, 1984) and does not depend on the availability of a complete three-dimensional structure. Since the covalent connections between regular secondary-structure elements can often also be unambiguously determined, this NMR information may be helpful in tracing the electron-density map for the determination of the corresponding X-ray crystal structure (Kallen *et al.*, 1991). Furthermore, since identical global architectures are usually found for corresponding crystal and solution structures of globular proteins, NMR structures have been used to solve the corresponding crystal structure by molecular replacement (*e.g.* Braun *et al.*, 1989). Considering the ease with which high-quality X-ray data are obtained nowadays once suitable crystals are available, it remains to be seen whether this kind of combined use of data obtained with the two methods will also play a role in the future.

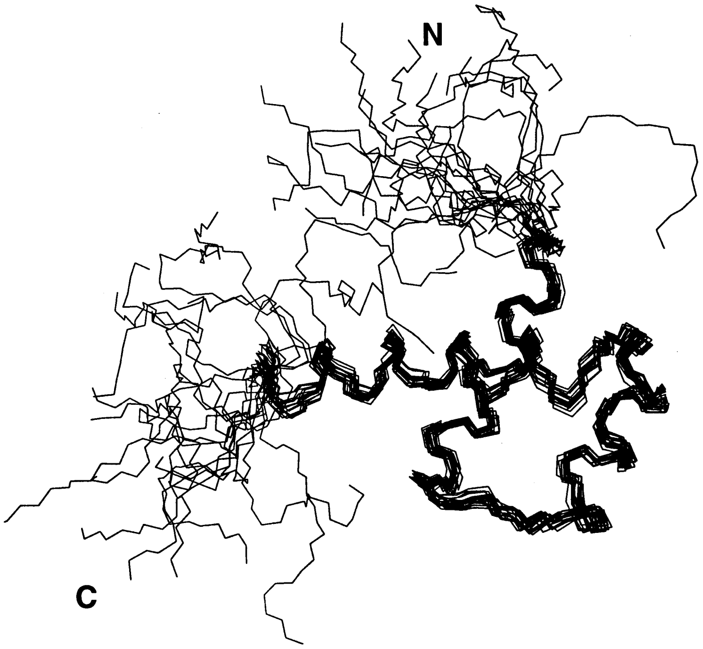


Figure 19.7.2.3
Polypeptide backbone of 19 energy-refined conformers selected to represent the NMR solution structure of the *Antennapedia* homeodomain. From residues 7 to 59, the structure is well defined by the NMR data. The chain-terminal segments 0–6 and 60–67 are disordered, and additional NMR studies showed that these terminal segments behave like 'flexible tails'. [Drawing prepared with the atomic coordinates from Qian *et al.* (1989).]

19.7.4. NMR studies of solvation in solution

In NMR structures, the location of hydration water molecules is determined by the observation of NOEs between water protons and hydrogen atoms of the polypeptide chain. In contrast to the observation of hydration water in X-ray crystal structures, this information is not routinely collected within the scope of a standard protocol for NMR structure determination (Fig. 19.7.2.1), but requires additional experiments (Otting *et al.*, 1991). Because the NOE decreases with the sixth power of the ^{1}H–^{1}H distance, only water molecules in a first hydration layer are typically observed. The NOE intensity is further related to a correlation function describing the stochastic modulation of the dipole–dipole coupling between the interacting protons, which may be governed either by the Brownian rotational tumbling of the hydrated protein molecule or by interruption of the dipolar interaction through translational diffusion of the interacting spins, whichever is faster (Otting *et al.*, 1991). Interior hydration waters are typically observed in identical locations in corresponding crystal and solution structures, but NMR provides additional information on molecular mobility. Completely buried hydration water molecules have thus been found to exchange with the bulk solvent at rates corresponding to millisecond residence times in the protein hydration sites, and measurements with nuclear magnetic relaxation dispersion also revealed exchange rates on the microsecond to nanosecond timescale (Denisov *et al.*, 1996). For surface hydration water, the lifetimes in the hydration sites are typically even shorter, and NMR measures the average duration of these 'visits' (Otting *et al.*, 1991;

in an alternative analysis of the NMR data, surface hydration has been characterized by reduced diffusion rates of the water, without specifying individual hydration sites; Brüschweiler & Wright, 1994). Diffraction experiments, on the other hand, probe the total fraction of time that a water molecule spends in a particular hydration site, but they are insensitive to the residence time at that site on any particular visit. Furthermore, while hydration water molecules in protein crystals are observed in discrete surface sites that are not blocked by direct protein–protein contacts, the entire surface of a protein in aqueous solution is covered with water molecules. Overall, the NMR view of hydration in solution, which has also been rationalized with long-time molecular-dynamics simulations (*e.g.* Billeter *et al.*, 1996; Brunne *et al.*, 1993), is largely complementary to crystallographic data on hydration.

19.7.5. NMR studies of rate processes and conformational equilibria in three-dimensional macromolecular structures

Similar to the aforementioned hydration studies, information on intramolecular rate processes in macromolecular structures cannot usually be obtained from the standard protocol for NMR structure determination (Fig. 19.7.2.1), but results from additional experiments. The complementarity of such NMR information to crystallographic data is well illustrated by the 'ring flips' of phenylalanine and tyrosine (Wüthrich, 1986). The observation of these ring-flipping motions in the basic pancreatic trypsin inhibitor (BPTI) (Wüthrich & Wagner, 1975) was a genuine surprise for the following reasons. In the refined X-ray crystal structure of BPTI, the aromatic rings of phenylalanine and tyrosine are among the side chains with the smallest temperature factors. For each ring, the relative values of the *B* factors increase toward the periphery, so that the largest positional uncertainty is indicated for carbon atom 4 on the symmetry axis through the $C^{\beta}-C^{1}$ bond, rather than for the carbon atoms 2, 3, 5 and 6 (Fig. 19.7.5.1), which undergo extensive movements during the ring flips. Theoretical studies then showed that the crystallographic *B* factors sample multiple rotation states about the $C^{\alpha}-C^{\beta}$ bond, whereas the ring flips about the $C^{\beta}-C^{1}$ bond seen by NMR are very rapid 180° rotations connecting two indistinguishable equilibrium orientations of the ring. The *B* factors do not manifest these rotational motions because the populations of all non-equilibrium rotational states about the $C^{\beta}-C^{1}$ bond are vanishingly small. The ring-flip phenomenon is now a well established feature of globular proteins, manifesting ubiquitous low-frequency internal motions with activation energies of 60–100 kJ mol^{-1}, amplitudes of $\gtrsim 1.0$ Å and activation volumes of about 50 Å^3 (Wagner, 1980), and involving concerted displacement of numerous groups of atoms (Fig. 19.7.5.1).

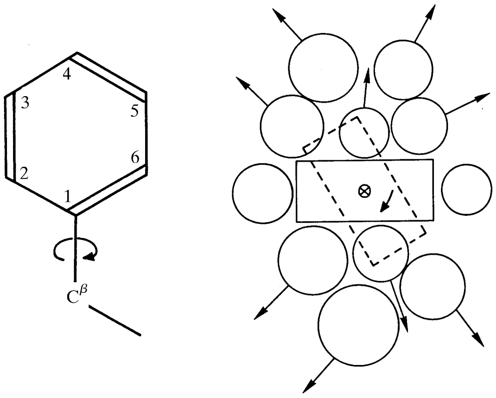


Figure 19.7.5.1
180° ring flips of tyrosine and phenylalanine about the $C^{\beta}-C^{1}$ bond. On the left, the atom numbering is given and the χ^2 rotation axis is identified with an arrow. The drawing on the right presents a view along the $C^{\beta}-C^{1}$ bond of a flipping ring in the interior of a protein, where the broken lines indicate a transient orientation of the ring plane during the flip. The circles represent atom groups near the ring, and arrows indicate movements of atom groups during the ring flip (Wüthrich, 1986).

References

Banci, L., Bertini, I., Cremonini, M. A., Gori-Savellini, G., Luchinat, C., Wüthrich, K. & Güntert, P. (1998). *PSEUDYANA for NMR structure calculation of paramagnetic metalloproteins using torsion angle molecular dynamics. J. Biomol. NMR*, **12**, 553–557.

Bax, A. & Grzesiek, S. (1993). *Methodological advances in protein NMR. Acc. Chem. Res.* **26**, 131–138.

Billeter, M. (1992). *Comparison of protein structures determined by NMR in solution and by X-ray diffraction in single-crystals. Q. Rev. Biophys.* **25**, 325–377.

Billeter, M., Braun, W. & Wüthrich, K. (1982). *Sequential resonance assignments in protein ^1H nuclear magnetic resonance spectra: computation of sterically allowed proton–proton distances and statistical analysis of proton–proton distances in single crystal protein conformations. J. Mol. Biol.* **155**, 321–346.

Billeter, M., Güntert, P., Luginbühl, P. & Wüthrich, K. (1996). *Hydration and DNA recognition by homeodomains. Cell*, **85**, 1057–1065.

Braun, W., Epp, O., Wüthrich, K. & Huber, R. (1989). *Solution of the phase problem in the X-ray diffraction method for proteins with the nuclear magnetic resonance solution structure as initial model. J. Mol. Biol.* **206**, 669–676.

Brünger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1986). *Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints. Application to crambin. Proc. Natl Acad. Sci. USA*, **83**, 3801–3805.

Brunne, R. M., Liepinsh, E., Otting, G., Wüthrich, K. & van Gunsteren, W. F. (1993). *Hydration of proteins. A comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations. J. Mol. Biol.* **231**, 1040–1048.

Brüschweiler, R. & Wright, P. E. (1994). *Water self-diffusion model for protein–water NMR cross relaxation. Chem. Phys. Lett.* **229**, 75–81.

Cavanagh, J., Fairbrother, W. J., Palmer, A. G. III & Skelton, N. J. (1996). *Protein NMR Spectroscopy, Principles and Practice*. New York: Academic Press.

Denisov, V. P., Peters, J., Hörlein, H. D. & Halle, B. (1996). *Using buried water molecules to explore the energy landscape of proteins. Nature Struct. Biol.* **3**, 505–509.

Dubs, A., Wagner, G. & Wüthrich, K. (1979). *Individual assignments of amide proton resonances in the proton NMR spectrum of the basic pancreatic trypsin inhibitor. Biochim. Biophys. Acta*, **577**, 177–194.

Ernst, R. R., Bodenhausen, G. & Wokaun, A. (1987). *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*. Oxford: Clarendon Press.

Güntert, P., Berndt, K. D. & Wüthrich, K. (1993). *The program ASNO for computer-supported collection of NOE upper distance constraints as input for protein structure determination. J. Biomol. NMR*, **3**, 601–606.

Güntert, P., Mumenthaler, C. & Wüthrich, K. (1997). *Torsion angle dynamics for NMR structure calculation with the new program DYANA. J. Mol. Biol.* **273**, 283–298.

Havel, T. F. & Wüthrich, K. (1985). *An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution. J. Mol. Biol.* **182**, 281–294.

Kallen, J., Spitzfaden, C., Zurini, M. G. M., Wider, G., Widmer, H., Wüthrich, K. & Walkinshaw, M. D. (1991). *Structure of human cyclophilin and its binding site for cyclosporin A determined by X-ray crystallography and NMR spectroscopy. Nature (London)*, **353**, 276–279.

Kay, L. E. & Gardner, K. H. (1997). *Solution NMR spectroscopy beyond 25 kDa. Curr. Opin. Struct. Biol.* **7**, 722–731.

Otting, G., Liepinsh, E. & Wüthrich, K. (1991). *Protein hydration in aqueous solution. Science*, **254**, 974–980.

Otting, G., Liepinsh, E. & Wüthrich, K. (1993). *Disulfide bond isomerization in BPTI and BPTI(G36S): an NMR study of correlated mobility in proteins. Biochemistry*, **32**, 3571–3582.

Peng, J. W. & Wagner, G. (1992). *Mapping of spectral density functions using heteronuclear NMR relaxation measurements. J. Magn. Reson.* **98**, 308–332.

Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. (1997). *Attenuated* $>T_2$ *relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc. Natl Acad. Sci. USA*, **94**, 12366–12371.

Qian, Y. Q., Billeter, M., Otting, G., Müller, M., Gehring, W. J. & Wüthrich, K. (1989). *The structure of the Antennapedia homeodomain determined by NMR spectroscopy in solution: comparison with prokaryotic repressors. Cell*, **59**, 573–580.

Riek, R., Wider, G., Pervushin, K. & Wüthrich, K. (1999). *Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. Proc. Natl Acad. Sci. USA*, **96**, 4918–4923.

Salzmann, M., Pervushin, K., Wider, G., Senn, H. & Wüthrich, K. (1998). *TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. Proc. Natl Acad. Sci. USA*, **95**, 13585–13590.

Tjandra, N. & Bax, A. (1997). *Direct measurement of distances and angles in biomolecules by NMR in dilute liquid crystalline medium. Science*, **278**, 1111–1114.

Wagner, G. (1980). *Activation volumes for the rotational motion of interior aromatic rings in globular proteins determined by high resolution* ^1H *NMR at variable pressure. FEBS Lett.* **112**, 280–284.

Wagner, G. & Wüthrich, K. (1982). *Sequential resonance assignments in protein* ^1H *nuclear magnetic resonance spectra: basic pancreatic trypsin inhibitor. J. Mol. Biol.* **155**, 347–366.

Wishart, D. S., Sykes, B. D. & Richards, F. M. (1991). *Relationship between nuclear-magnetic-resonance chemical shift and protein secondary structure. J. Mol. Biol.* **222**, 311–333.

Wüthrich, K. (1976). *NMR in Biological Research: Peptides and Proteins*. Amsterdam: North Holland.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. New York: Wiley.

Wüthrich, K. (1989). *Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science*, **243**, 45–50.

Wüthrich, K. (1995). *NMR – this other method for protein and nucleic acid structure determination. Acta Cryst.* D**51**, 249–270.

Wüthrich, K., Billeter, M. & Braun, W. (1984). *Polypeptide secondary structure determination by nuclear magnetic resonance observation of short proton–proton distances. J. Mol. Biol.* **180**, 715–740.

Wüthrich, K. & Wagner, G. (1975). *NMR investigations of the dynamics of the aromatic amino acid residues in the basic pancreatic trypsin inhibitor. FEBS Lett.* **50**, 265–268.

International Tables for Crystallography (2012). Vol. F, Chapter 19.8, pp. 620–623.

Chapter 19.8. Use of *SPIDER* and *SPIRE* in image reconstruction

A. Leith, W. Baxter and J. Frank

19.8.1. Introduction

SPIDER (System for Processing Image Data in Electron microscopy and Related fields) was one of the earliest image-processing packages for single-particle reconstruction and electron tomography. It was first released in 1978, and has been continually improved and updated since then (Frank *et al.*, 1981, 1996; Baxter *et al.*, 2007; Shaikh *et al.*, 2008; Yang *et al.*, 2007). It is currently available as free open-source code under Gnu Public License (GPL).

In its most common usage, *SPIDER* allows a researcher to create a three-dimensional (3D) reconstruction of a macromolecule from a collection of transmission electron micrographs, interpreted as projections showing the molecule in different orientations. Using cryo-electron microscopy these methods, pioneered by Joachim Frank at the Wadsworth Center in Albany (see Frank, 2006), have recently allowed researchers to determine the 3D structures of several macromolecular complexes at near-atomic resolution (~3.8–4.5 Å; Zhou, 2008). These methods provide a way of studying the structure of such complexes in a more natural environment than is possible using X-ray crystallography and of studying conformational changes of molecular machines as they perform their work (Saibil, 2000).

One way in which *SPIDER* differs from most other software for single-particle reconstruction [*Imagic* (van Heel *et al.*, 1996; Chapter 19.9), *Frealign* (Grigorieff, 2007), *EMAN2* (Tang *et al.*, 2007; Chapter 19.10), *SPARX* (Hohn *et al.*, 2007) and *Xmipp* (Sorzano *et al.*, 2004)] is that it provides alternative methods for many procedures and thus can serve as a developmental platform for testing different approaches to single-particle reconstruction. For instance, alternative alignment operations and statistical analysis operations can be combined inside scripts to investigate different approaches to overcoming molecular heterogeneity.

The prime reasons for the continued success of *SPIDER* are its validated reputation for providing the highest-resolution reconstructions, its ease of installation and its comprehensive documentation, which is constantly being updated (Shaik *et al.*, 2008). *SPIDER*'s procedure language makes it easy to use for controlling the flow of a complex train of operations and addressing problems related to heterogeneity of particles. With the Quick Start Guide and tutorials as an introduction, a new user can quickly acquire competence in the procedure language with only a few hours' introduction.

The *SPIDER* system consists of six major components:

(i) *SPIDER*: the Fortran program with a command interpreter recognizing *SPIDER* commands and procedure calls.

(ii) *SPIDER* procedures: text files containing *SPIDER* commands, parameters, and script-specific operations for conditional execution and looping (see Section 19.8.3).

(iii) *Web*: a graphical user interface (GUI) for use in Linux and OS X (Frank *et al.*, 1996). *Web* is available in two different versions, the original X-Window version written in C and a newer Java version.

(iv) *SPIDER* reconstruction engine (*SPIRE*), which represents a metastructure that enables procedure files, file numbering and directories required for a single-particle reconstruction project to be managed from a simple GUI (see Section 19.8.4; Baxter *et al.*, 2007).

(v) *PubSub*: a set of Perl programs that enable *SPIDER* procedures to be run in parallel on a computer cluster.

(vi) *SPIDER* documentation: an extensive collection of more than 800 HTML-based documents.

The system is available in versions for Linux, AIX and OS X. Both source code and precompiled binaries for popular platforms are provided. *SPIDER* is available for free download as a GPL Open Source distribution. Its documentation is available under a Creative Commons Attribution 2.5 Licence.

19.8.2. Basic philosophy of single-particle reconstruction

'Single-particle reconstruction' is the term used for the reconstruction of a biological macromolecule from images of a specimen in which the molecule exists in many 'copies' in the form of single isolated particles, *i.e.* with no contact with neighbouring molecules. Since there is no need for crystallization, there is in principle no restriction on the kinds of macromolecules that can be reconstructed, except that they must be above a critical size required for accurate alignment. Combined with cryogenic electron microscopy (cryo-EM), the method is capable of visualizing molecules in their native states. Although the method was originally conceived for a homogeneous population, the introduction of powerful classification techniques is now allowing heterogeneous populations to be disentangled and represented by a series of reconstructions which, suitably ordered, may reflect the development of a system of interacting molecules (a molecular machine) over time.

Each particle image is interpreted as a noisy projection of the three-dimensional Coulomb potential representing the molecule, which is, for practical purposes, identical to the electron-density distribution rendered by X-ray crystallography. The noise (signal-to-noise ratio $\simeq 0.1$) is due mainly to the low exposure required to avoid radiation damage. For reconstruction, all projections must be placed in a common coordinate frame. Since, unlike the case in electron tomography, the angles are initially unknown, the most challenging and computationally intensive task in single-particle reconstruction is the determination of particle orientations, usually done in an iterative manner with increasing angular resolution. In the following, it is assumed that a reference density map of a closely related molecular complex is already available. For instance, a map of an empty ribosome may be used as reference for a data set obtained from a ribosome complexed with EF-G, mRNA and tRNA. For *ab initio* reconstructions of an unknown structure, random-conical and common-lines techniques are available (see Shaikh *et al.*, 2008).

Owing to the oscillatory behaviour of the contrast transfer function (CTF), an entire defocus series must be collected in order to cover the whole range of information in Fourier space. For each micrograph, the defocus is determined by computing the power spectrum and matching it with the CTF. Two strategies are in use in the field: CTF correction is done either at the stage of the raw micrograph, by phase flipping and pooling all data for

Copyright © 2012 International Union of Crystallography

reconstruction, or by pooling data into groups with a narrow defocus range, which are then separately reconstructed and combined at the very end using CTF correction by Wiener filtering (Penczek *et al.*, 1997). Although both strategies can be readily realized in *SPIDER*, the one most thoroughly tested and proven is the latter, sometimes referred to as defocus group reconstruction.

For a step-by-step description of the processing path, we refer to the flow diagram in Fig. 19.8.2.1. After defocus determination, particles are selected and aligned to a set of reference projections using rotational and translational cross-correlation. Eulerian angles yielding the highest cross-correlation are assigned. Next, the data are pooled into defocus groups, defined by the range of allowed defocus variation within each group. The range needs to be small enough to avoid resolution loss. Reconstruction of each defocus group yields a preliminary density map. These density maps are merged by CTF correction *via* Wiener filtering, yielding an initial CTF-corrected reconstruction. This reconstruction is then used as the reference in the first angular refinement of defocus group reconstructions, which follows the same principle as the initial alignment. Several cycles of angular refinement are used with progressively smaller angular steps, from the initial ~15° to 0.5° or even 0.2°. From a certain point on, comparisons are only done within a narrow cone of the previous Eulerian angle assignment for a given projection. The progress of the refinement is monitored by computing the Fourier shell correlation at each stage and by following the statistics of angle reassignment, watching for evidence of stability.

For supervised classification, not one but two three-dimensional references are used in the projection alignment (Fig. 19.8.2.1), and the assignment of the class and set of Eulerian angles is done according to the parameters of the reference projection producing the highest cross-correlation coefficient. For example, ribosome data may be split on the basis of inter-subunit ratchet rotation, by presenting 'ratcheted' and 'non-ratcheted' ribosomes as references.

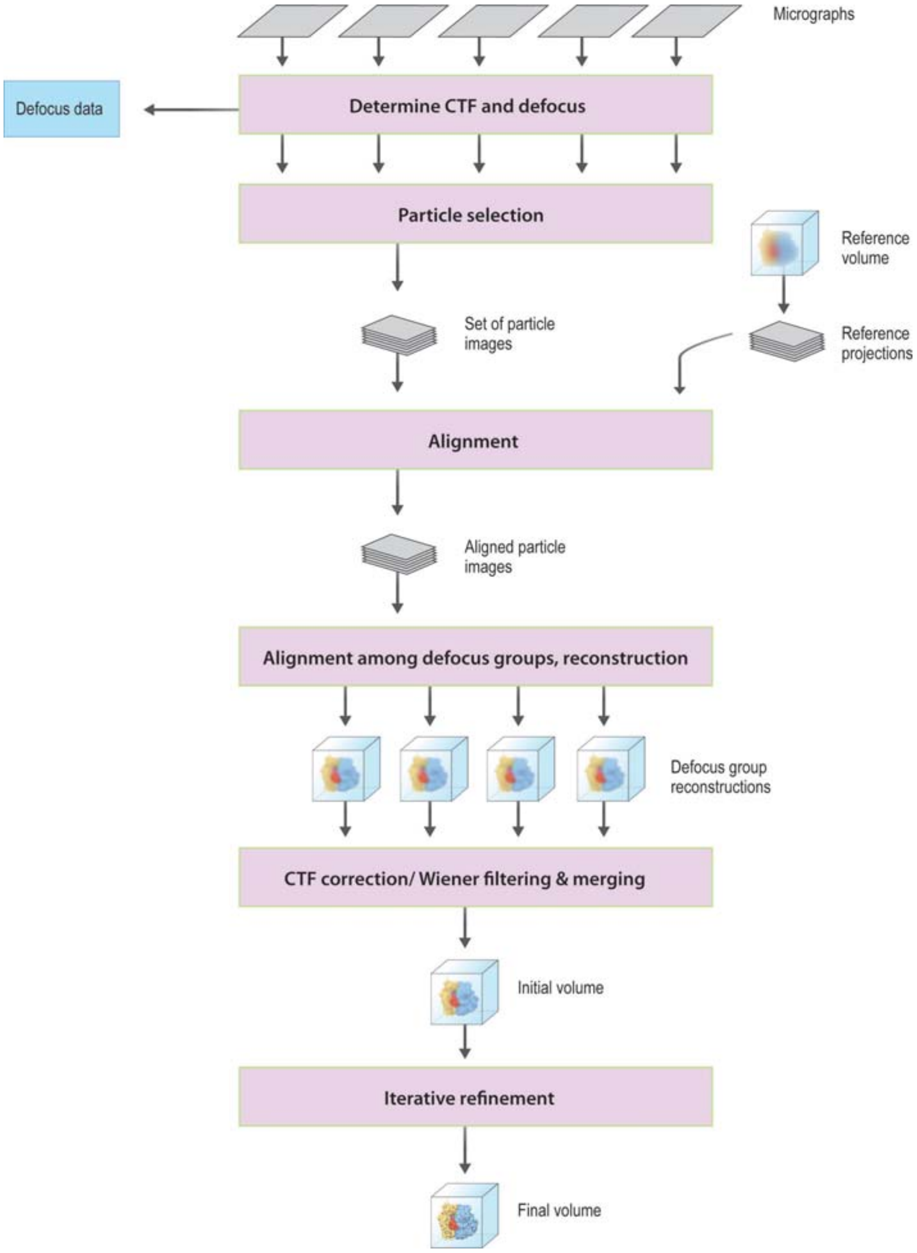


Figure 19.8.2.1
Flow diagram of reference-based single-particle reconstruction by defocus groups. For details, see text.

19.8.3. Implementation of single-particle reconstruction in *SPIDER*

The implementation of the reconstruction strategy presented in the flow diagram (Fig. 19.8.2.1) makes use of the modular features of *SPIDER* and its hierarchical calling structure (Fig. 19.8.3.1). The computational task of the entire project (level 4, or uppermost in the hierarchy) can be visualized as being composed of building blocks (level 3), each representing a subtask, such as image alignment, defocus determination, angular refinement *etc.* Each of these building blocks is realized by either a set of elementary commands (level 1) or procedure files containing such commands (level 2). In the reconstruction task, examples of procedures are `deffsc.spi` and `ctf.spi`, which compute the resolution of each defocus group and apply CTF correction, respectively. On the uppermost project level, all data file names, directories and values of parameters must be specified, while the input for lower-level procedures can be variables whose value is not known until the time of execution.

19.8.3.1. Level 1: elementary commands

Commands in *SPIDER* consist of a short two-to-three letter operation mnemonic and a variable sub-option mnemonic. For example, `BP CG` invokes a program performing back-projection reconstruction using a conjugate gradient algorithm. In the interactive on-line mode, commands are entered at a prompt and the system responds with a series of subsequent prompts for more input specific to the operation the command invokes. In the off-line batch mode, relevant in the context of time-consuming reconstructions, both commands and subsequent inputs are placed in a batch or procedure file in the expected order for successive execution.

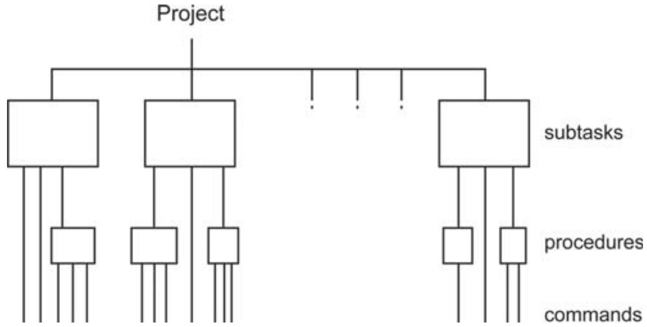
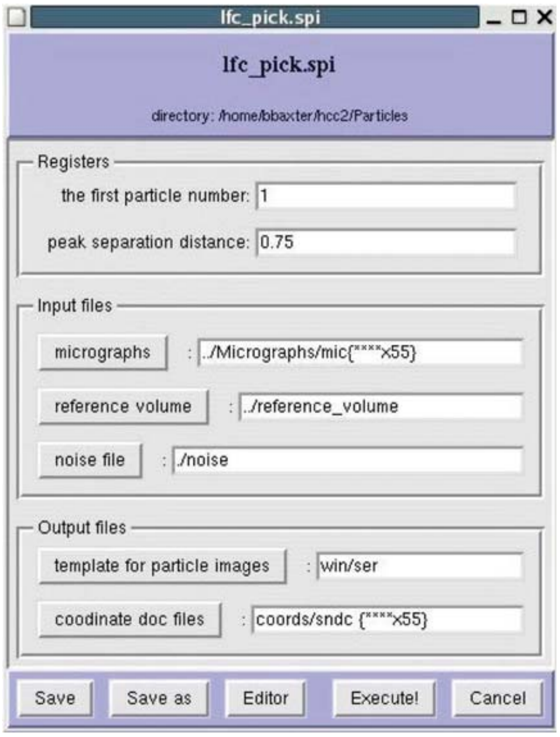


Figure 19.8.3.1
Schematic representation of the *SPIDER* hierarchy of functional units and calling levels.



lfc_pick.spi

directory: /home/bbaxter/hcc2/Particles

Registers

the first particle number: 1

peak separation distance: 0.75

Input files

micrographs : ../Micrographs/mic{****x55}

reference volume : ../reference_volume

noise file : ./noise

Output files

template for particle images : win/ser

coodinate doc files : coords/sndc {****x55}

Figure 19.8.4.1
A *SPIRE* form, displaying the parameters, input files and output files of a *SPIDER* procedure file. The user can edit the values in the white entry boxes or browse for files (*e.g.* by clicking on the 'noise file' button). The changes can either be saved to update the procedure file or executed in *SPIDER* immediately ('Save' and 'Execute' buttons).

There are over 400 different computational operations and sub-operations available in *SPIDER*, which can be grouped into the following categories:

(i) *File handling*: copying, format converting, montaging, stacking.
(ii) *Real-space image manipulations*: alignment, centre-of-gravity computation, contouring, contrast adjustment, edge detection, filtering, histogram computation, masking, mirroring, montaging, peak search, rotation, scaling, segmentation, thresholding, windowing.
(iii) *Fourier-space image manipulations*: autocorrelation, cross-correlation, CTF correction, Fourier filtering, Fourier interpolation, Fourier transformation, resolution determination.
(iv) *Statistical analysis*: averaging, classification of image sets, multivariate data analysis.
(v) *Three-dimensional reconstruction*: algebraic reconstruction technique (ART), constrained simultaneous iterative reconstruction technique (SIRT), projection, SIRT, weighted back-projection.

19.8.3.2. Level 2: procedures

The modular design of *SPIDER* allows incremental development and flexibility by allowing the user to write scripts in the form of procedure files. Procedures containing sequences of commands can be executed just like single commands. Procedures have their own name space, can pass arguments to other procedures and can request input during an interactive *SPIDER* session.

A single-particle reconstruction project requires the application of many complex operations on perhaps several hundred thousand images, in a 30- to 40-step process that involves branching and decision making. To handle these tasks, the scripting facilities of *SPIDER* include commands for conditional expressions, iteration over numbered file names and document manipulation.

SPIDER stores and retrieves numerical results in document files. These are keyed sequential text files containing a table of values, and are used to permanently store items such as image shift coordinates, rotation angles and particle numbers. These values can be retrieved

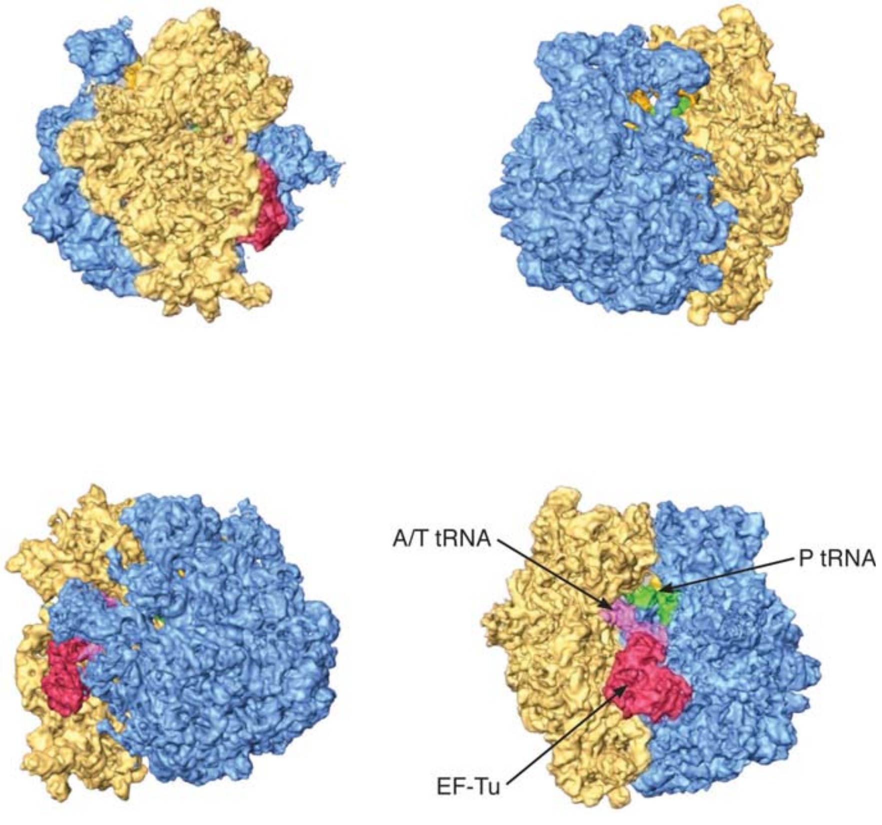


Figure 19.8.5.1
Reconstruction of the *E. coli* ribosome bound with the Phe-tRNA·EF-Tu·GDP ternary complex at 6.7 Å resolution, obtained by the protocol outlined here (LeBarron *et al.*, 2008).

by different procedures or even subsequent *SPIDER* sessions.

There are over 30 document and scripting operations in *SPIDER*:

(i) *Document operations*: retrieve values, select items, sort document, store values.

(ii) *Scripting operations*: assign numerical variables, assign string variables, capture values from *SPIDER* operations, check numerical values, carry out conditional execution and looping, evaluate mathematical expressions, make file enquiries, pass values to/from procedures, set global switches.

19.8.4. Implementation of single-particle reconstruction in *SPIRE*

SPIRE (the *SPIDER* reconstruction engine) is an interactive GUI for managing a *SPIDER* reconstruction (Baxter *et al.*, 2007). A project of single-particle reconstruction may involve dozens of procedure files, each of which may contain hundreds of individual *SPIDER* operations. Procedures usually read *SPIDER* input files (images, volumes or text document files), carry out some processing and write out new *SPIDER* files. The outputs of one procedure are the inputs of the next, and perhaps of other procedures further down the processing stream. Hundreds of thousands of files may be generated during the course of a project.

SPIRE was developed to manage the numerous procedures and data files involved in a *SPIDER* project. As each procedure is executed, *SPIRE* keeps track of the output files generated and adds them to a project database. Users can run the same procedure several times, changing input parameters, and each 'run' is given a unique identification number by *SPIRE*. *SPIDER* is executed in a controlled environment by *SPIRE*, in which error conditions can be trapped and appropriate messages displayed to the user. Successful runs are added to the project database. *SPIRE* provides a graphical interface to these tasks, allowing the user to make changes to the inputs and outputs of a procedure in a displayed form (Fig. 19.8.4.1). Project-wide parameters used by multiple procedures are also graphically viewable and editable. The project database has a spreadsheet-like interface, which can display *SPIDER* images, volumes and stack files in Jweb, as well as text documents in a user-specified editor. Procedure files can be edited, new procedures may be added to a project, and the way a procedure is displayed can be changed.

Projects are represented in *SPIRE* by configuration files. These are text files that use XML tags to specify content, such as the order of procedure files and the directory structure of a project. *SPIRE* provides a configuration editor to manipulate configurations, and thus the overall structure of a *SPIDER* project. If there is a laboratory database system, *SPIRE* provides some capability to upload project results using SQL commands. Finally, *SPIRE* allows the user to write the project file to disk in HTML format, so that a web browser can be used to find out when batch files were run and which data files were generated during the project. *SPIRE* is written in the Python programming language; for graphics it uses the Tkinter module (the interface to the Tk library).

19.8.5. Conclusion

SPIDER was designed to allow high versatility in the design of program flow, so the protocol presented here and in the article by Shaikh *et al.* (2008), which has been used successfully in many applications (for a recent example, see Fig. 19.8.5.1), can be readily changed to meet the challenges of a new project by users who have received a modest amount of training. The same goes for the implementation of alternative strategies, such as CTF correction of raw data, which is beyond the scope of this chapter.

This work was funded by HHMI and grants from NCRR/NIH P41 RR01219 (to JF), P01 GM064692 (to Robert M. Glaeser) and NIH R01 GM150601 (to AL). The authors thank Lila Iino-Rubenstein for assistance with the illustrations.

References

Baxter, W. T., Leith, A. & Frank, J. (2007). *SPIRE: the SPIDER reconstruction engine. J. Struct. Biol.* **157**, 56–63.

Frank, J. (2006). Editor. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies – Visualization of Biological Molecules in Their Native State.* New York: Oxford University Press.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M. & Leith, A. (1996). *SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol.* **116**, 190–199.

Frank, J., Shimkin, B. & Dowse, B. H. (1981). *SPIDER: a modular software system for electron image processing. Ultramicroscopy*, **6**, 343–358.

Grigorieff, N. (2007). *FREALIGN: high-resolution refinement of single particle structures. J. Struct. Biol.* **157**, 117–125.

Heel, M. van, Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. (1996). *A new generation of the IMAGIC image processing system. J. Struct. Biol.* **116**, 17–24.

Hohn, M., Tang, G., Goodyear, G., Baldwin, P. R., Huang, Z., Penczek, P. A., Yang, C., Glaeser, R. M., Adams, P. D. & Ludtke, S. J. (2007). *SPARX, a new environment for Cryo-EM image processing. J. Struct. Biol.* **157**, 47–55.

LeBarron, J., Grassucci, R. A., Shaikh, T. R., Baxter, W. T., Sengupta, J. & Frank, J. (2008). *Exploration of parameters in cryo-EM leading to an improved density map of the E. coli ribosome. J. Struct. Biol.* **164**, 24–32.

Penczek, P. A., Zhu, J., Schroeder, R. & Frank, J. (1997). *Three-dimensional reconstruction with contrast transfer function compensation from defocus series. Scanning Microsc.* **11**, 147–154.

Saibil, H. R. (2000). *Conformational changes studied by cryo-electron microscopy. Nat. Struct. Biol.* **7**, 711–714.

Shaikh, T. R., Gao, H., Baxter, W. T., Asturias, F. J., Boisset, N., Leith, A. & Frank, J. (2008). *SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs. Nat. Protoc.* **3**. 1941–1974.

Sorzano, C. O., Marabini, R., Velázquez-Muriel, J., Bilbao-Castro, J. R., Scheres, S. H., Carazo, J. M. & Pascual-Montano, A. (2004). *XMIPP: a new generation of an open-source image processing package for electron microscopy. J. Struct. Biol.* **148**, 194–204.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. (2007). *EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol.* **157**, 38–46.

Yang, C., Penczek, P. A., Leith, A., Asturias, F. J., Ng, E. G., Glaeser, R. M. & Frank, J. (2007). *The parallelization of SPIDER on distributed-memory computers using MPI. J. Struct. Biol.* **157**, 240–249.

Zhou, Z. H. (2008). *Towards atomic resolution structural determination by single-particle cryo-electron microscopy. Curr. Opin. Struct. Biol.* **18**, 218–228.

references

International Tables for Crystallography (2012). Vol. F, Chapter 19.9, pp. 624–628.

Chapter 19.9. Four-dimensional cryo-electron microscopy at quasi-atomic resolution: *IMAGIC 4D*

M. van Heel, R. Portugal, A. Rohou, C. Linnemayr, C. Bebeacua, R. Schmidt, T. Grant and M. Schatz

19.9.1. Introduction

During the 1950s and 1960s the electron microscope became a routine research instrument. Many biological structures we now take for granted, such as the ribosome (Chemistry Nobel Prize 2009, see www.nobel.se), were first discovered using the electron microscope. An avalanche of such discoveries emerged after the appropriate preparation techniques for biological specimens were developed, such as the contrasting of molecular complexes with heavy-metal salts (Brenner & Horne, 1959). This dry ‘negative-stain’ approach, however, can also create structural deformations of the sample, hampering its detailed analysis. The introduction of the vitreous water (‘vitreous ice’) specimen preparation technique by Jacques Dubochet and co-workers (Adrian *et al.*, 1984) represented a huge improvement in structural preservation for biological specimens in the harsh vacuum environment of the electron microscope. Notwithstanding the contributions made to the cryo-electron microscopy (cryo-EM) field by the data-processing methodology and software – the subject of this chapter – it was the vitreous ice specimen preparation approach that marked the future success of single-particle cryo-EM. For all practical purposes, vitreous ice is water and the first electron images of virus particles trapped in their native solution state were spectacular (Adrian *et al.*, 1984). The vitreous ice preparations can trap the molecular complexes in full action, in different conformational states, and may represent a window into the detailed functioning of the complexes in their natural environment.

Structural analysis by single-particle cryo-EM is therefore very different to that carried out by traditional X-ray crystallography. With the latter technique, the data collected are diffraction intensities, necessarily reflecting already *averaged* contributions from all unit cells in the crystal. Any information relating to the behaviour of individual molecules is thus lost during data collection. Single-particle cryo-EM, in contrast, is unique in that it preserves the identity of each individual molecular image registered in whatever structural state it happens to be. The marked difference with X-ray crystallography is that the information about each individual molecule or complex is explicitly available in digital form, albeit strongly deteriorated by noise. This also means that in cryo-EM all averaging takes place in the computer; the computational requirements of cryo-EM therefore exceed those of X-ray crystallography by orders of magnitude.

The intrinsic heterogeneity of cryo-EM samples was originally known as the curse of cryo-EM because it degrades the quality of the resulting average structures. Many structures simply would not refine to a reasonable resolution [for reviews of the single-particle methods see van Heel *et al.* (2000) and Frank (2006)]. In recent years, however, heterogeneity has been seen more as a blessing, since, with the appropriate advanced data-processing technology, it is now possible to separate the different conformations of the complex into individual three-dimensional structures. Each of these three-dimensional structures may reveal a different state of the biological complex, as was first shown with a mixture of functional states of the *E. coli* 70S ribosome in complex RF3 (Klaholz *et al.*, 2004). For recent reviews on the challenges posed by heterogeneous biological samples, see Leschziner & Nogales (2007) and Spahn & Penczek (2009). These fascinating new developments build upon almost four decades of advances in software instrumentation, which started in the mid-1960s (DeRosier & Klug, 1968). We here review recent developments of the *IMAGIC* system aimed at optimally processing large heterogeneous ‘four-dimensional’ (4D) cryo-EM data sets to high resolution.

19.9.2. The *IMAGIC* software system

The *IMAGIC* software system (van Heel & Keegstra, 1981; van Heel *et al.*, 1996) is the result of over 30 years of continuous developments, including various major redesign phases. It was one of the first software packages in electron microscopy and many now-standard procedures were pioneered in *IMAGIC*, including surface rendering, multivariate statistical analysis (MSA), automatic classification, projection matching, angular reconstitution, automatic particle selection *etc*. The basic file format of *IMAGIC* consists of a header file, to hold image parameters and processing information, and a separate file holding all actual image data. This ‘stack’ format can hold millions of individual molecular images and all *IMAGIC* programs are designed to loop over all the images in the stack (van Heel & Keegstra, 1981). Users are not burdened with the task of formulating loops over the two-dimensional images needing the same treatment.

The *IMAGIC* philosophy is to create specialized high-level programs for all more complicated tasks. High-level programs allow much better interactive guidance for the user than do image-processing scripts (which are operating-system dependent). A unique feature of all *IMAGIC* programs is that all user interaction is concentrated in a ‘user interaction block’ (UIB). Whether interactively processing images or creating a large ‘batch job’ for later processing, the user always communicates with the program’s UIB with all interactive guidance available (each question comes with its own interactive help). The *IMAGIC* system is available for the most popular current operating systems including Linux, Windows XP/Vista/7 and Mac OS-X.

19.9.3. *IMAGIC* ‘4D’ processing/data format

Recently, another major overhaul of the system was undertaken in order to better handle three-dimensional structural heterogeneity. The relevant *IMAGIC* programs can now loop over entire three-dimensional volumes (and not just two-dimensional images) allowing programs to tackle problems at a much higher level of complexity. About 25 years ago, a radical redesign of the system allowed small two-dimensional images to reside ‘in core’ (the main memory of the computer) rather than on disk, facilitating all alignment algorithms. The new redesign allows whole

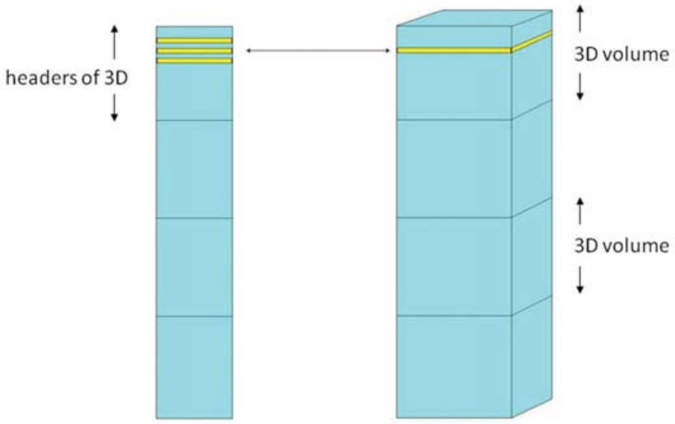


Figure 19.9.3.1
IMAGIC 4D file format. The 4D file format is not very different to the original file format with one header record per 2D image in a stack (van Heel & Keegstra, 1981). The main difference is that specific header locations now indicate the additional organization of a sequence of 2D images in the stack belonging together in a 3D volume. Substantial changes were required in the software, however, to make programs loop over the 3D volumes in one 4D *IMAGIC* file in the way that the programs looped over the 2D images in a stack. For full format details, see www.imagescience.de.

three-dimensional volumes to remain in core and to be manipulated efficiently by single subroutine calls without I/O (input/output) overhead. The file format for '4D' processing has hardly needed any updating other than defining some new header parameters. The reason is that a stack of three-dimensional (3D) volumes (a '4D' data format) is still just a stack of two-dimensional (2D) images, the traditional *IMAGIC* data format (Fig. 19.9.3.1). This '4D' upgrade, however, did require a radical overhaul of all programs, since all relevant programs needed an extra loop over a set of 3D volumes. At the same time, these 3D programs have all become much faster, more compact and more easily maintained because they rely on the new in-core 3D libraries and are now freed of all excess I/O calls.

19.9.4. Software parallelization

In cryo-EM all averaging and information extraction from all individual molecular images takes place *in silico*; the computational requirements are thus huge. Computers are never fast enough for the most demanding single-particle approaches. Over the last 15 years much emphasis has been placed on the parallelization of critical *IMAGIC* code (van Heel *et al.*, 2000), mainly using a message passing interface (MPI; Gropp *et al.*, 1994) to take advantage of modern 'cluster' computing environments. Other software packages have since followed the same parallelization path (Smith & Carragher, 2008). The *IMAGIC* software is implemented such that the same code will run on all machines from a single-CPU notebook computer up to large cluster systems with hundreds of CPUs.

Recently, a 'GPU' library (see www.nvidia.com/object/cuda_home.html) has been implemented in *IMAGIC* to exploit the properties of cheap graphic processors. However, for most 'standard library' operations, like 2D fast Fourier transforms, which can be almost directly linked to existing programs, most time gained in GPU processing is still lost in transporting data to and from the limited memory of the GPUs. The question is whether the time invested in software development for exploiting specific properties of any specific parallel computing hardware will pay off in the longer term for the specific needs of cryo-EM. In the case of the GPUs, the amount of memory typically available per GPU core is rather low.

19.9.5. Full 2D (parallel) astigmatic contrast transfer function correction

The *IMAGIC* contrast transfer function (CTF) estimation and correction programs (van Heel *et al.*, 2000) have recently been upgraded to work fully in two dimensions, enabling the accurate detection of all CTF parameters including astigmatism. This has been carried out in order to accomplish two goals. Firstly, the programs are now capable of operating on an entire data set of charge-coupled device (CCD) images or patches of micrographs. This now allows the use of the parallel MSA programs (see below) to classify sets of amplitude spectra (and create class averages thereof) prior to the precise determination of the defocus and astigmatism parameters. This enables the simple and largely automatic determination of CTF parameters, and subsequent CTF correction *via* phase flipping of entire data-set stacks. Secondly, the ability to accurately detect even extreme levels of astigmatism enables the use of highly astigmatic images, unlocking extremely close-to-focus defoci in order to push the achievable resolution. These programs have been used for the processing of a highly astigmatic data set of the *Limulus polyphemus* hemocyanin, resulting in a ~4 Å reconstruction from only 15 000 raw molecular images, corresponding to 60 000 asymmetric units for this $C2$ point-group symmetry structure (Grant *et al.*, 2011).

19.9.6. Parallel automatic particle picking

Parallel processing is of increasing importance for collecting the large raw data sets required for 4D cryo-EM. If we are interested in a structure that represents 1% of all molecular complexes, we need to increase the size of the data set 100-fold in order to achieve the same resolution we had for a monodisperse data set. This implies that we need to be able to rapidly process data sets of the order of 1 Tbyte in size. One of the fastest automatic particle-selection procedures is still one of the oldest ones, based on the local variance (van Heel, 1982) in the raw micrographs or CCD images. It is applied immediately after the full-data-set CTF correction discussed above. Very important for the calculation of the variance image is the choice of the frequency range used for discriminating the presence of a particle with respect to the background. The sensitivity of the approach is approximately as good as that of particle searching using the cross-correlation function (CCF) (Saxton & Frank, 1977). However, like all CCF alignments (see below), the CCF particle picking requires templates and may tend to bias the particle selection towards the references (Boekema *et al.*, 1986; Stewart & Grigorieff, 2004). Modulation-image particle picking is a variant of the original variance image detection (van Heel, 1982) which is based on the local standard deviation rather than the local variance and avoids squaring of amplitudes. This new algorithm is often the method of choice for an unbiased automatic particle selection. Once a reliable first 3D structure has been calculated, the particle-picking program can then look for all possible views in all possible orientations in an extensive stack of input images (the massive calculations involved exploit MPI parallelization).

19.9.7. Parallel multi-reference alignments and reference bias

Large multi-reference alignment (MRA) schemes (van Heel & Stöffler-Meilicke, 1985) have been accelerated by some orders of magnitude with the introduction of MPI parallelization on computer clusters (van Heel *et al.*, 2000), and MRA remains the largest source of CPU needs in single-particle cryo-EM. However, with the recent focus on discriminating between different functional states, the issue of reference bias resurfaces. With the renewed emphasis on avoiding reference bias (Boekema *et al.*, 1986; Stewart & Grigorieff, 2004), 'alignment by classification' of centred molecular images (Dube *et al.*, 1993) has taken on a new emphasis. An especially useful variant is the rotational alignment of the data set with respect to the main symmetry-related eigenvector. This alignment does not bias the data set with respect to specific references, yet it concentrates much of the variance in the data set into the lower eigenimages (van Heel *et al.*, 2009) and thus yields better class averages while using the same total number of eigenvectors. Such new unbiased approaches are still in full development.

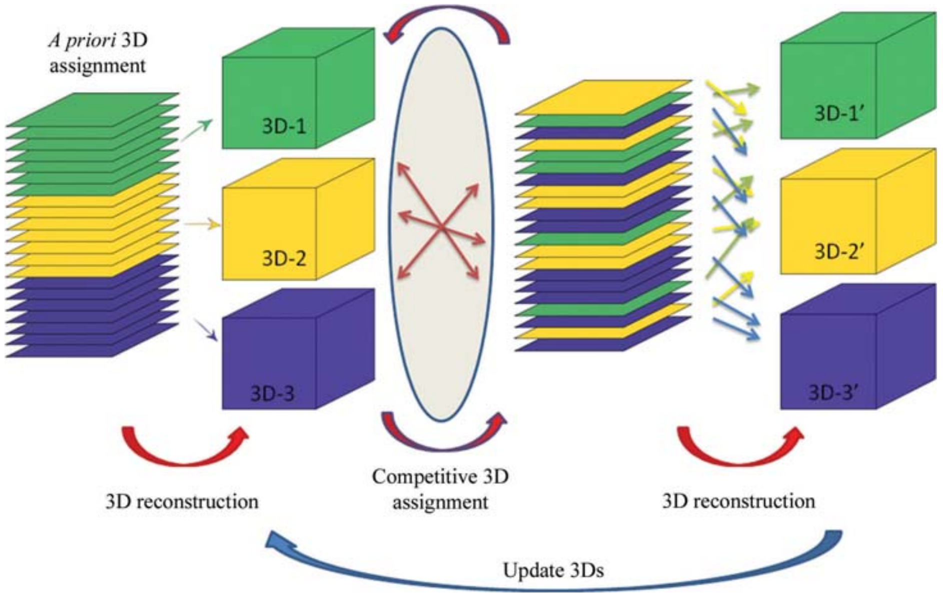


Figure 19.9.7.1
One starts with the (*a priori*) assignment of the images in a stack (original images or class averages) to a number of 3D volumes (named 3D-1 to 3D-3). This assignment can be of various origins: 2D MSA manifold separation; 3D MSA classification; or even a random-number generator, depending on the strategy pursued. Once different 3Ds have been generated, various 'competitive' 3D membership assignment schemes can be applied to the 2D images in the stack. A classical one is multi-reference alignment with respect to re-projections of the different 3D volumes. Another approach is the multi-anchor set Euler angle assignment by 'angular reconstitution' as described in the text. The resulting new 3Ds (named 3D-1′ to 3D-3′) take the place of the earlier 3Ds in this generic iterative refinement scheme. The iterations are stopped once convergence criteria like the Fourier shell correlation (FSC) resolution of the 3Ds reach stability (Harauz & van Heel, 1986; van Heel & Schatz, 2005).

MRAs are now used as a tool in 4D analysis by generating MRA references from a set of 3D structures rather than just a single one (Klaholz *et al.*, 2004; Spahn & Penczek, 2009). Although not different from the 2D versions of MRA, the new *IMAGIC 4D* programs make it simpler to explicitly track a '3D membership' indicating to which 3D volume the best alignment was achieved. The overall administration of the 4D procedures is greatly facilitated, especially during iterative refinement rounds (Fig. 19.9.7.1)

19.9.8. MSA and its parallelization

Multivariate statistical analysis (MSA) was originally introduced to electron microscopy to handle different views of a macromolecule (van Heel & Frank, 1981; Borland & van Heel, 1990; van Heel *et al.*, 2000; Frank, 2006; for an extensive review, see van Heel *et al.*, 2009). MSA approaches are the techniques of choice to structure complex, heterogeneous data sets. Today, a wealth of information on MSA techniques is readily available on the Internet. The highly efficient *IMAGIC* MSA programs are specifically optimized for large data sets: the CPU time required for calculating the main eigenvectors is directly proportional to the total size of the data set. The naive use of standard MSA libraries in most EM software leads to eigenvector/eigenvalue calculations that are proportional to the *square* of the number of images in a data set, making it impossible to process more than a few thousand images (van Heel *et al.*, 2009).

In the first 4D cryo-EM paper revealing different functional states of the same structure (Klaholz *et al.*, 2004), the MSA separation of the data set into the two states was achieved by a manually supported separation of the two conformers in eigenvector space, each representing a different functional state of the 70S ribosome in complex with RF3. Each functional state of a complex represents a 'manifold' in factor space, and one seeks to separate the different manifolds in an unambiguous way, possibly by weighting the eigenvectors (van Heel, 1984; Elad *et al.*, 2008). An alternative is to use the MSA programs to explicitly analyse large sets of 3D volumes in the way one traditionally analyses large sets of 2D images (see van Heel *et al.*, 2009).

The ever-increasing size of the data sets posed a new problem: the MSA processing of large data sets containing, say, a million images (totalling, say, ~1 Tbyte) takes months to process on a single CPU. An MPI parallelization project was undertaken to speed up the eigenvector/eigenvalue calculations. This required parallelization of not only the pure computations but also the I/O operations, since the algorithm reads the large input data sets many times (~30 times) in order to iteratively refine towards a stable answer, leading to I/O bottlenecks (van Heel *et al.*, 2009). Like all other programs mentioned in this chapter, the parallel MSA program is now part of the normal *IMAGIC* distribution.

19.9.9. Handling multiple 3D reconstructions in parallel

A new parallel 3D reconstruction program was created based on the earlier ex-core program implementing the 'exact filter' algorithm (Harauz & van Heel, 1986). With all 3D reconstructions now organized in in-core subroutines, the new program is organized to loop over many 3D volumes and to generate whole series of 3D reconstructions using new options. It is an ideal tool for generating (large) 4D data sets as needed, for example, for the 4D MSA approaches. It is a perfect match for the 4D MRA (above) and the 4D angular reconstitution programs (Fig. 19.9.7.1).

The new program, apart from the exact filter algorithm, now also features a novel 3D deconvolution algorithm based on the idea that one can predict the 3D point-spread function (PSF) for any 3D reconstruction geometry. The correct 3D reconstruction is then calculated by the 3D deconvolution of the unfiltered 3D reconstruction with the PSF. This 3D deconvolution algorithm scales proportionally to the number of projections used to generate a 3D reconstruction (the exact filter algorithm scales as the *square* of the number of projections) and yields results of the same overall quality as the classical exact filter algorithm (Harauz & van Heel, 1986).

19.9.10. Angular reconstitution 4D refinements

With the angular reconstitution approach (van Heel, 1987; van Heel *et al.*, 2000) one typically assigns a Euler angle orientation to a 2D image by finding its best overall sinogram correlation peaks with respect to an 'anchor set' of 2D projections from one single 3D structure. The *IMAGIC* 'Euler' program, which performs the angular reconstitution orientational search, has now been extended to loop over multiple anchor sets and has thus been upgraded to a 4D level. As a consequence, the Euler program is now capable of competitively choosing the best anchor set among anchor sets generated from different 3D structures. The program can thus now also be used for refining towards multiple 3D structures, as was already the case with the MRA program in *IMAGIC*.

19.9.11. Discussion

The potential benefits of 4D cryo-EM at atomic resolution are obvious. With some further developments, 4D cryo-EM may become the structural-biology method *par excellence*. The growing importance of software in cryo-EM has been the subject of various editorials and special issues of journals (Carragher & Smith, 1996; Smith & Carragher, 2008). The processing of 4D data sets, however, requires solving a large set of 3D structures simultaneously and that is necessarily a complex matter. The researcher will have to understand the basic principles of the approach and its pitfalls, and will thus have to go through a steep learning curve. The basic methodology (and the use of *IMAGIC*) is taught in courses such as the EMBO course for 3D cryo-EM, or the Brazil school for single-particle cryo-EM (www.single-particles.org/school/). The *IMAGIC 4D* system is aimed at accompanying the user at the appropriate level of complexity, without forcing the user to become an expert script programmer. Throughout the *IMAGIC 4D* system, the software will directly understand the 4D frame of thinking of the user.

There is a growing understanding in the field that continuity of development, maintenance and support of the software for cryo-EM is an important issue for the success of a project. One needs to safeguard the human investment in mastering the learning curve, and one needs continuity of support over the duration of a (long-term) project. Software continuity is at least as important as it is to continue maintaining electron microscopes for data collection in cryo-EM. An example of a successful community effort to consolidate software instrumentation in the field of X-ray crystallography is the maintenance of the *CCP4* suite of programs (Collaborative Computational Project, Number 4, 1994). One must realize, however, that this model cannot be transferred directly to the cryo-EM field because each of the handful of packages in use in single-particle cryo-EM is at least as complex as the *CCP4* system, largely due to the necessity of parallelization.

One further remark about using scripts for running existing software: scripts tend to lead a life of their own and are often exchanged among users; combinations of scripts are sometimes even distributed as a 'new' package. However, scripts are operating-system dependent, and any change in the original programs requires updating of all associated scripts in all different operating systems. The lack of 'backward' compatibility with existing scripts actually impedes the development of the original software, since many users may refuse to update the core software to avoid having to modify their (borrowed) scripts. Creating wrappers (Python) to interface with original programs has a similar effect of possibly hampering software evolution.

Another compatibility issue is that of the formats used in cryo-EM. The original 2D stack format of *IMAGIC* (van Heel & Keegstra, 1981) is one of the richest formats in use in cryo-EM and therefore it is possible to convert from all file formats in use in cryo-EM into *IMAGIC* without loss of information. The popular *EM2EM* conversion program (www.imagescience.de) for converting all cryo-EM formats into each other is an *IMAGIC* program that converts formats without any unnecessary loss of header data. The popular *CCP4* volume data format (Collaborative Computational Project, Number 4, 1994) has been adopted as a standard for the current EM databases (Smith & Carragher, 2008). This format is a 3D format, with one header record for the full 3D volume. It can also hold a single 2D image, but it cannot hold a stack of images, each with its own header record, as is required for single-particle cryo-EM. For 4D cryo-EM applications a 4D format will be required for the structure databases. There is some urgency for the cryo-EM and X-ray communities to agree on a more elaborate common density format for our structure databases.

Atomic resolution structures (~3 Å) have hitherto been elucidated mainly by X-ray crystallography, where the biological molecules are confined to the rigidity of a 3D crystal. Single-particle cryo-EM gives a direct window into the solution, revealing a plethora of different views, of different complexes, in different functional states, albeit at a resolution level still somewhat inferior to that achieved by X-ray crystallography and NMR spectroscopy. Single-particle cryo-EM techniques are now approaching resolution levels previously only achievable by X-ray crystallography. For a better understanding of biological processes, it is essential to see the sequence of conformational changes and interactions that molecules undergo during their functional cycle. A primary challenge in structural biology today is to generate '4D movies' of biological complexes at (quasi)-atomic resolution. The new *IMAGIC 4D* software has been tailored for such 4D analysis at atomic resolution.

We acknowledge recent financial support from the EU/NOE (grant No. NOE-PE0748), from the Dutch Ministry of Economic Affairs (Cyttron project No. BIBCR_PX0948) and from the BBSRC (grant No. BB/G015236/1). We thank all users of the *IMAGIC* system for constructive feedback and past contributors for their code and ideas. We also acknowledge the use of the *Eclipse/SVN* software suite for helping to keep the *IMAGIC* system coherent with contributors operating from three different continents.

References

Adrian, M., Dubochet, J., Lepault, J. & McDowall, A. W. (1984). *Cryo-electron microscopy of viruses. Nature (London)*, **308** 32–36.

Boekema, E. J., Berden, J. A. & van Heel, M. G. (1986). *Structure of mitochondrial F1-ATPase studied by electron microscopy and image processing. Biochim. Biophys. Acta*, **851**, 353–360.

Borland, L. & van Heel, M. (1990). *Classification of image data in conjugate representation spaces. J. Opt. Soc. Am. A*, **7**, 601–610.

Brenner, S. & Horne, R. W. (1959). *A negative staining method for high resolution electron microscopy of viruses. Biochim. Biophys. Acta*, **34**, 103–110.

Carragher, B. & Smith, P. R. (1996). *Advances in computational image processing for microscopy. J. Struct. Biol.* **116**, 2–8.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

DeRosier, D. J. & Klug, A. (1968). *Reconstruction of three-dimensional structures from electron micrographs. Nature (London)*, **217** 130–134.

Dube, P., Tavares, P., Lurz, R. & van Heel, M. (1993). *The portal protein of bacteriophage SPP1: a DNA pump with 13-fold symmetry. EMBO J.* **12**, 1303–1309.

Elad, N., Clare, D. K., Saibil, H. R. & Orlova, E. V. (2008). *Detection and separation of heterogeneity in molecular complexes by statistical analysis of their two-dimensional projections. J. Struct. Biol.* **162**, 108–120.

Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford University Press.

Grant, T., van Duinen, G., Patwardhan, A. & van Heel, M. (2011). *Exploiting astigmatism in single particle cryo electron microscopy*. Submitted.

Gropp, W., Lusk, E. & Skjellum, A. (1994). *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. Scientific and Engineering Computation Series. Cambridge, MA: MIT Press.

Harauz, G. & van Heel, M. (1986). *Exact filters for general geometry three dimensional reconstruction. Optik*, **73** 146–156.

Heel, M. van (1982). *Detection of objects in quantum noise-limited images. Ultramicroscopy*, **7**, 331–342.

Heel, M. van (1984). *Multivariate statistical classification of noisy images (randomly oriented biological macromolecules). Ultramicroscopy*, **13**, 165–183.

Heel, M. van (1987). *Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy*, **21**, 111–124.

Heel, M. van & Frank, J. (1981). *Use of multivariate statistics in analyzing the images of biological macromolecules. Ultramicroscopy*, **6**, 187–194.

Heel, M. van, Gowen, B., Matadeen, R., Orlova, E. V., Finn, R., Pape, T., Cohen, D., Stark, H., Schmidt, R., Schatz, M. & Patwardhan, A. (2000). *Single-particle electron cryo-microscopy: towards atomic resolution. Q. Rev. Biophys.* **33**, 307–369.

Heel, M. van, Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. (1996). *A new generation of the IMAGIC image processing system. J. Struct. Biol.* **116**, 17–24.

Heel, M. van & Keegstra, W. (1981). *IMAGIC: a fast flexible and friendly image analysis software system. Ultramicroscopy*, **7**, 113–130.

Heel, M. van, Portugal, R. & Schatz, M. (2009). *Multivariate statistical analysis in single particle (cryo) electron microscopy*. In *An Electronic Text Book: Electron Microscopy in Life Science, 3D-EM Network of Excellence*, edited by A. Verkley & E. V. Orlova. London: 3D-EM Network of Excellence.

Heel, M. van & Schatz, M. (2005). *Fourier shell correlation threshold criteria. J. Struct. Biol.* **151**, 250–262.

Heel, M. van & Stöffler-Meilicke, M. (1985). *Characteristic views of E. coli and B. stearothermophilus 30S ribosomal subunits in the electron microscope. EMBO J.* **4**, 2389–2395.

Klaholz, B. P., Myasnikov, A. G. & van Heel, M. (2004). *Release factor 3 seen on the ribosome during termination of protein synthesis. Nature (London)*, **427**, 862–865.

Leschziner, A. E. & Nogales, E. (2007). *Visualizing flexibility at molecular resolution: analysis of heterogeneity in single-particle electron microscopy reconstructions. Annu. Rev. Biophys. Biomol. Struct.* **36**, 43–62.

Saxton, W. O. & Frank, J. (1977). *Motif detection in quantum noise-limited electron micrographs by cross-correlation. Ultramicroscopy*, **2** 219–227.

Smith, R. & Carragher, B. (2008). *Software tools for molecular microscopy. J. Struct. Biol.* **163**, 224–228.

Spahn, C. M. T. & Penczek, P. A. (2009). *Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. Curr. Opin. Struct. Biol.* **19**, 1–9.

Stewart, A. & Grigorieff, N. (2004). *Noise bias in the refinement of structures derived from single particles. Ultramicroscopy*, **102**, 67–84.

International Tables for Crystallography (2012). Vol. F, Chapter 19.10, pp. 629–632.

Chapter 19.10. Single-particle reconstruction with *EMAN*

S. Ludtke

19.10.1. Introduction

Single-particle reconstruction is a technique for determining the three-dimensional structure of identical nanoscale objects without requiring crystallization (Frank, 2006). Typical targets are large proteins or macromolecular assemblies, with ~200 kDa regarded as the lower size limit for this technique due to high image noise levels. Typical non-viral targets are in the size range from 500 kDa to 5 MDa, with icosahedral viruses extending to hundreds of megadaltons. Several examples of structures being reconstructed to ~4 Å resolution using this technique have been published (Ludtke *et al.*, 2008; Jiang *et al.*, 2008; Zhang *et al.*, 2008; Yu *et al.*, 2008) and subnanometre resolution is frequently achieved, but in negative stain, or with unusually difficult specimens, resolutions may still be limited to 12–30 Å.

19.10.2. Overview of *EMAN*

EMAN is a suite of scientific image-processing tools with a particular focus on cryogenic electron microscopy (cryo-EM) and single-particle reconstruction. The first stable version was released in 1999 and it has undergone continuous development since that time. Since roughly 2005, development has focused on *EMAN2*, with a completely redesigned modular image-processing library, a workflow system, and a new openGL-based graphical user interface (GUI). *EMAN2* supports virtually all image file formats used in the cryo-EM community, as well as various conventions for particle-orientation specification. The image-processing library is completely modular, so new algorithms can be made available throughout the system without any modification of end-user programs. Metadata, which refer to both experimental and derived information describing the image data, are archived using an embedded database system. These metadata can be harvested to produce information for manuscripts or for deposition to centralized databases such as the EMDB (http://www.emdatabank.org), operated by the PDBe (Europe) and RCSB PDB (USA).

EMAN2 has a tiered structure. The GUI interface and workflow represent the top of the hierarchy. Next are individual command-line programs, providing the user with a greater level of control and a wide range of utility functions. Below this is the Python (http://www.python.org) layer, for interactive use or high-level programming. Finally, we have the C++ core image-processing library, which handles all computationally intensive image processing. Full documentation is provided through the *EMAN* wiki (http://blake.bcm.tmc.edu/eman/eman2).

19.10.2.1. GUI layer and workflow

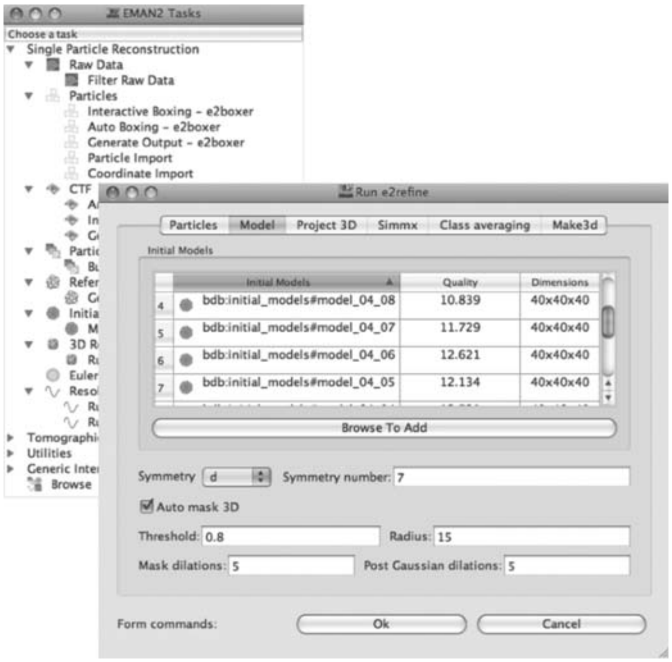
The workflow interface (Fig. 19.10.2.1) is designed to guide the end user through entire processes, while storing a complete record of all processing completed during any given task. While this interface is gradually being expanded with new tasks, such as tomographic particle averaging, for the purposes of this discussion we will focus on the single-particle reconstruction task, which is broken down into a number of discrete subtasks as discussed in detail in Section 19.10.3. The workflow is designed so that users wishing to perform only a subset of any task can easily import data into any subtask, proceed through additional subtasks, then export the results at any point. There are future plans to integrate other cryoEM software into the workflow interface to enable users to try alternative algorithms for specific subtasks. The workflow is accessed using either `e2workflow.py` or `e2desktop.py`.

The GUI interface (Fig. 19.10.2.2) includes a number of visualization modules for displaying individual two-dimensional images, stacks of two-dimensional images, three-dimensional volumetric models, two-dimensional plots, three-dimensional plots and symmetry information. These modules are used by the interactive file/database browser available in the workflow or *via* the program `e2display.py`, but may also be used directly from the Python interface.

19.10.2.2. Command-line programs

EMAN2 (Tang *et al.*, 2007) currently contains ~60 command-line programs for specific tasks, plus an additional ~60 programs as part of the related *SPARX* project (Hohn *et al.*, 2007), included with *EMAN2* releases. High-level programs such as `e2refine2d.py`, `e2initialmodel.py`, `e2eotest.py` and `e2refine.py` are essentially wrapper scripts which call the lower-level Python scripts to complete specific tasks.

In addition, there are a number of general-purpose utility programs, such as `e2proc2d.py` for two-dimensional image processing, `e2proc3d.py` for three-dimensional image processing and `e2bdb.py` for access to the embedded database. Both `e2proc2d.py` and `e2proc3d.py` take similar arguments and



		Initial Models	Quality	Dimensions
4	●	bdb:initial_models#model_04_08	10.839	40x40x40
5	●	bdb:initial_models#model_04_07	11.729	40x40x40
6	●	bdb:initial_models#model_04_06	12.621	40x40x40
7	●	bdb:initial_models#model_04_05	12.134	40x40x40

Symmetry: d
Symmetry number: 7
☑ Auto mask 3D
Threshold: 0.8
Radius: 15
Mask dilations: 5
Post Gaussian dilations: 5

Figure 19.10.2.1
The overall workflow interface, and the dialogue for the three-dimensional refinement subtask.

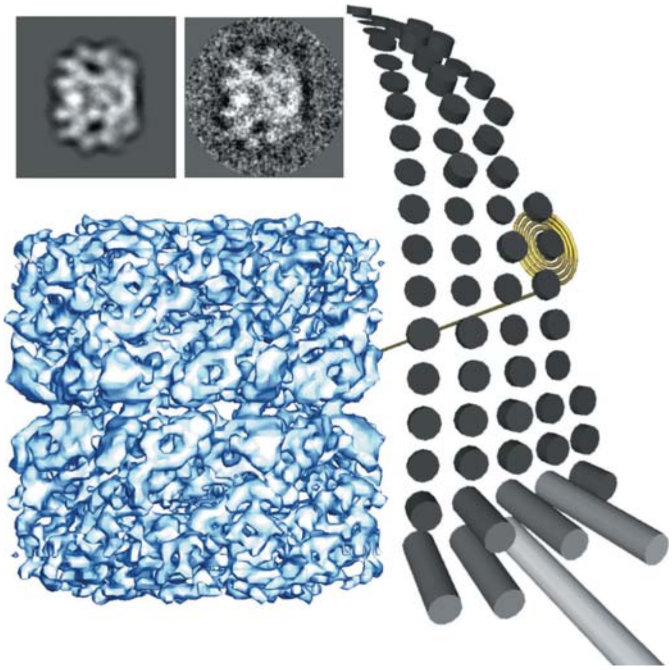


Figure 19.10.2.2
A representation of the Euler GUI tool. The three-dimensional reconstruction is shown (lower left), along with a projection and class average for a particular orientation identified in the asymmetric triangle on the right. Each cylinder in the asymmetric triangle represents a projection direction, with its height representing the number of particles found in that orientation.

permit application of any of the ~170 modular image-processing algorithms in the core library *via* the `--process` option. These include filters, masks, mathematical operations and a range of other processing algorithms. The `e2help.py` program provides detailed documentation for all of the modular algorithms. All *EMAN2* programs are able to read automatically any of the supported file formats, and will write to any supported file format based on file extension or explicit specification of output format. For example, the command to convert a three-dimensional volume from HDF format to MRC format is simply `e2proc3d.py input.hdf output.mrc`.

19.10.2.3. Python wrapper

Python is an easy-to-learn scripting language in wide use in both scientific and non-scientific disciplines. It provides both an interactive prompt as well as a programming interface, and is easily extensible using higher-performance languages such as C++. *EMAN2* was designed such that all command-line programs and the entire GUI interface are written in Python. Thus, any of the distributed programs in *EMAN2* can be modified by experienced end users without a full *EMAN2* compilation environment. In addition, the program `e2.py` offers an *iPython*-based (http://ipython.scipy.org) interactive prompt which gives full access to all of the library functions and GUI tools of *EMAN2*. An introduction to this interface can be found at http://blake.bcm.edu/emanwiki/Eman2ProgQuickstart.

19.10.2.4. C++

The C++ interface contains a modular system, so algorithms of various types can be trivially made available to the entire workflow system. The C++ and Python interfaces were intentionally designed to be as similar to each other as possible, so prototypes of new algorithms can be developed in Python and then later ported to C++. See http://blake.bcm.edu/emanwiki/Eman2CProgQuickstart for an introduction.

Table 19.10.2.1
Listing of *EMAN2* supported file formats and whether each has read (R) and/or write (W) support

BDB refers to the *EMAN2* embedded database. DM2/3 are Gatan Digital Micrograph formats. LST files are text files from *EMAN1*.

Format	R or W support	Format	R or W support
BDB	R/W	TIFF	R/W
HDF5	R/W	PGM	R/W
MRC	R/W	PNG	R/W
IMAGIC	R/W	JPEG	W
SPIDER	R/W	LST	R/W
PIF	R/W	*AMIRA*	R/W
DM3	R	*XPLOR*	W
DM2	R	VTK	W
EM	R/W	*FITS*	R/W
ICOS	R/W	SAL	R

19.10.2.5. Cross-platform support

All of the major computing platforms are supported, including Linux workstations and clusters, Mac OSX and Windows. Our primary development platforms are Linux and OSX. While we are committed to full support of *EMAN2* on Windows, we have encountered idiosyncratic behaviour on some specific machines running Vista which have defied explanation, but we continue to strive for full support.

19.10.2.6. Parallel processing

The three-dimensional refinement process can be extremely computationally intensive. While a refinement of a small ~1 MDa particle with some symmetry to 15 Å resolution may be accomplished on a single workstation in a few hours, a project like an asymmetric reconstruction to 4 Å resolution could easily require hundreds of thousands of CPU hours. Updated details of the modular parallelism strategy of *EMAN* and its GPGPU (general purpose graphics processing unit) methodology can be found in the wiki at http://blake.bcm.edu/emanwiki/EMAN2/Parallel.

19.10.2.7. File formats and other conventions

EMAN2 supports all cryo-EM file formats for which specifications were available, in addition to its embedded database storage mechanism. While any *EMAN2* program can read and write images in any supported format, we have adopted HDF5 as an interchange format and the internal database BDB for use during workflow operations. These two permit arbitrary metadata to be associated with each image, unlike the standard cryo-EM formats. Table 19.10.2.1 contains a list of the currently supported formats.

The other primary convention of concern to single-particle reconstruction is three-dimensional orientation specification. While *EMAN2* has its own convention, it can convert to and from the most common conventions in use in the cryo-EM community, including *MRC*, *SPIDER*, *IMAGIC*, quaternions and spin axis. *EMAN*'s own convention uses Z-X-Z' Euler angles named az, alt and phi, respectively.

19.10.3. Single-particle reconstruction

The single-particle reconstruction process is broken down into a sequence of subtasks in the *EMAN2* workflow. Each subtask is documented in the GUI, with a block of text and mouse-over popups. While the workflow is the method of choice for performing reconstructions, all of the individual subtasks can also be completed using direct command-line programs. The workflow is launched with the `e2workflow.py` program. For each stage, the alternative command-line program is also shown in parentheses.

19.10.3.1. Particle selection (`e2boxer.py`)

Locating images of individual particles in the micrograph/charge-coupled device (CCD) frame is arguably the most critical and labour-intensive step in single-particle processing. Owing to the high noise levels in typical cryo-EM images, combined with various types of contamination, correctly identifying all relevant particles within a micrograph is challenging (Zhu *et al.*, 2004). Despite many years of effort, no reliable and generally applicable algorithms for fully automated particle picking exist. The present approach of *EMAN2* is to streamline the process of semi-automated particle picking. Multiple CCD frames/micrographs can be opened simultaneously and are handled by the user as a group. At the time of writing, there is a single autopicking algorithm available, based on Woolford *et al.* (2007), but there are plans to expand this in the future.

19.10.3.2. Contrast transfer function/image evaluation (`e2ctf.py`)

EMAN1 contained a very accurate contrast transfer function (CTF) correction methodology, which was also very labour-intensive. While the mathematical basis for correction remains effectively unchanged, the correction methodology is now entirely automatic. The only substantial change to the mathematical formulation is that the spectroscopic profile of the background noise, which was originally characterized using four empirical parameters, is now specified as a data-derived non-parametric curve. In addition, this methodology permits the computation of a one-dimensional structure factor from sets of images without user intervention, a process which was quite challenging in *EMAN1* (Ludtke *et al.*, 1999). The automated process is supplemented by a fully featured GUI for manual image assessment.

19.10.3.3. Grouping particles into sets

Once images have been preprocessed, the user is given another, optional, opportunity to examine the images of individual particles, now in Wiener-filtered form, to identify any 'bad' particles which should be eliminated from further processing. In addition, several values are provided to the user which permit assessment of particles on a per-micrograph basis, such as the integrated signal-to-noise ratio, defocus and B factor. Particles from user-selected micrographs are then combined into a set for further processing.

19.10.3.4. Reference-free two-dimensional classification (`e2refine2d.py`)

This fully automated principal-component-analysis-based (Frank, 2006) process produces two-dimensional averages representative of the various particle views present in the raw images. As with the workflow as a whole, while the user can provide a large number of different options for this subtask, sensible default values are provided, which should produce good results in most cases. The two-dimensional classification process is quite similar to the corresponding process in *EMAN1* (Chen *et al.*, 2006), with subtle improvements to increase speed and accuracy.

The two-dimensional classification process generates class averages with improved signal-to-noise, characteristic of the different orientations and dynamic states of the particles. These can be used to help identify the particle's symmetry and any problems with preferred orientation or particle flexibility/heterogeneity.

19.10.3.5. Initial model generation (`e2initialmodel.py`)

There are a variety of opinions in the single-particle reconstruction community about how initial models should be generated. Briefly, the most widely used methods are: common lines with reference-free class averages (van Heel, 1987), random conical tilt (Radermacher *et al.*, 1987), orthogonal tilt (Leschziner & Nogales, 2006) and single-particle tomography with three-dimensional averaging (Walz *et al.*, 1997). While not the primary method we promote for traditional single-particle processing, the *EMAN2* workflow includes a task for three-dimensional alignment and averaging of tomographic subvolumes.

The primary method provided to generate initial models in *EMAN2* is based on the robustness of the iterative refinement strategy used for the final high-resolution refinement (Section 19.10.3.6). For any given three-dimensional structure, there are a small number of stable structures which can result from refinement, regardless of starting model. The correct model represents the global minimum in the energy space defined by requiring that projections of the reconstruction match the particle images.

To make use of this concept in *EMAN2*, we downsample the two-dimensional class averages (Section 19.10.3.4) and make a large number of completely random starting models, then quickly iteratively refine each. While the speedups used in this process make it more susceptible to local minima than the full high-resolution refinement method, it will generally still produce the correct model some fraction of the time, without having to resort to additional experiments. The resulting models are easily assessed by comparing the original class averages with projections of the model, so the best answer can be readily identified. The primary risk associated with this process is with respect to particles with structural heterogeneity in solution or those with strongly preferred orientations, in which case the standard refinement algorithm is not robust. In such cases, the tomographic method can provide further insights.

19.10.3.6. Refinement (`e2refine.py`)

Once particles have been prepared and an initial model has been produced, the next step is three-dimensional refinement. This process is largely unchanged from *EMAN1* (Ludtke *et al.*, 1999), apart from minor improvements designed to speed up the process and provide additional flexibility. As with two-dimensional refinement, there are a large number of options for the user to specify, but, where possible, sensible defaults are provided. Documentation for the various parameters is provided through the workflow interface.

19.10.4. Evaluating the reconstruction

Evaluating the final three-dimensional reconstruction remains one of the most problematic aspects of single-particle processing from a quantitative standpoint, and there have been numerous debates in the community over standards for model assessment. As the field is now beginning to achieve resolutions where protein side chains can be visualized, reliable new assessments based on methods from X-ray crystallography are emerging, but at lower resolutions robust assessment remains elusive.

19.10.4.1. Model accuracy

Before assessing the resolution of a reconstruction, the fundamental question of whether the model even qualitatively represents the original data must be addressed. Regardless of the reconstruction methodology used, the fundamental questions to be asked are whether computed projections of the reconstruction match both the raw particles and class averages, and whether all of the particle views are represented by the three-dimensional model in some orientation. The 'Eulers' option in the workflow provides a number of tools for making such model assessments both qualitatively and quantitatively.

19.10.4.2. Measures of resolution and resolvability

Resolution in single-particle processing is distinct from the related concept of resolvability. Resolvability is a measure of the level of detail visible in a model, in terms of the shortest separation distance over which two objects can be identified as being distinct. However, in structural biology, 'resolution' is a statement of the spatial frequency at which the noise level exceeds a threshold. It is important to recognize that a model with 4 Å resolution could be low-pass filtered to a resolvability of only 20 Å, and yet cryo-EM resolution measures would still (properly) show it to have 4 Å resolution. To help bring the resolvability in line with the resolution, it is typical to apply an appropriate filter to the three-dimensional reconstruction such that the resolvability is in reasonable agreement with the resolution, but there remains no consensus in the community over the optimal filter and/or filtration level appropriate for this task. *EMAN2* provides a signal-to-noise-ratio- and structure-factor-based technique, as well as two mechanisms for assessing the resolution of a reconstruction.

19.10.4.3. Model/noise bias

The final issue to consider in a single-particle reconstruction is the well known model/noise bias problem (Stewart & Grigorieff, 2004). With a traditional iterative refinement strategy, and very high noise levels in the raw particle images, it is possible to produce a reconstruction including features derived from the initial model or from systematic algorithmic artifacts which are not represented in the raw data. There are relatively few techniques for assessing this sort of bias, and each single-particle reconstruction package handles this issue differently. In *EMAN*, the use of iterative class averaging during the iterative refinement process permits this bias to be greatly reduced or eliminated, when used as suggested. Ensuring that the reference-free class averages agree well with projections of the reconstruction can at least place some limits on the extent of such artifacts.

The development of *EMAN2* is funded by NIH grant No. R01GM080139. I would like to thank David Woolford for his work on the figures, and Ben Bammes and Jesus Montoya for their comments.

References

Chen, D. H., Song, J. L., Chuang, D. T., Chiu, W. & Ludtke, S. J. (2006). *An expanded conformation of single-ring GroEL-GroES complex encapsulates an 86 kDa substrate. Structure*, **14**, 1711–1722.

Frank, J. (2006). *Three-dimensional electron microscopy of macromolecular assemblies: visualization of biological molecules in their native state*. In *Multivariate Data Analysis and Classification of Images*. Oxford University Press.

Heel, M. van (1987). *Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. Ultramicroscopy*, **21**, 111–123.

Hohn, M., Tang, G., Goodyear, G., Baldwin, P. R., Huang, Z., Penczek, P. A., Yang, C., Glaeser, R. M., Adams, P. D. & Ludtke, S. J. (2007). *SPARX, a new environment for cryo-EM image processing. J. Struct. Biol.* **157**, 47–55.

Jiang, W., Baker, M. L., Jakana, J., Weigele, P. R., King, J. & Chiu, W. (2008). *Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. Nature (London)*, **451**, 1130–1134.

Leschziner, A. E. & Nogales, E. (2006). *The orthogonal tilt reconstruction method: an approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. J. Struct. Biol.* **153**, 284–299.

Ludtke, S. J., Baker, M. L., Chen, D. H., Song, J. L., Chuang, D. T. & Chiu, W. (2008). *De novo backbone trace of GroEL from single particle electron cryomicroscopy. Structure*, **16**, 441–448.

Ludtke, S. J., Baldwin, P. R. & Chiu, W. (1999). *EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol.* **128**, 82–97.

Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. (1987). *Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. J. Microsc.* **146**, 113–136.

Stewart, A. & Grigorieff, N. (2004). *Noise bias in the refinement of structures derived from single particles. Ultramicroscopy*, **102**, 67–84.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I. & Ludtke, S. J. (2007). *EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol.* **157**, 38–46.

Walz, J., Typke, D., Nitsch, M., Koster, A. J., Hegerl, R. & Baumeister, W. (1997). *Electron tomography of single ice-embedded macromolecules: three-dimensional alignment and classification J. Struct. Biol.* **120**, 387–395.

Woolford, D., Ericksson, G., Rothnagel, R., Muller, D., Landsberg, M. J., Pantelic, R. S., McDowall, A., Pailthorpe, B., Young, P. R., Hankamer, B. & Banks, J. (2007). *SwarmPS: rapid, semi-automated single particle selection software. J. Struct. Biol.* **157**, 174–188.

Yu, X., Jin, L. & Zhou, Z. H. (2008). *3.88 Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. Nature (London)*, **453**, 415–419.

Zhang, X., Settembre, E., Xu, C., Dormitzer, P. R., Bellamy, R., Harrison, S. C. & Grigorieff, N. (2008). *Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction. Proc. Natl Acad. Sci. USA*, **105**, 1867–1872.

Zhu, Y., Carragher, B., Glaeser, R. M., Fellmann, D., Bajaj, C., Bern, M., Mouche, F., de Haas, F., Hall, R. J., Kriegman, D. J., Ludtke, S. J., Mallick, S. P., Penczek, P. A., Roseman, A. M., Sigworth, F. J., Volkmann, N. & Potter, C. S. (2004). *Automatic particle selection: results of a comparative study. J. Struct. Biol.* **145**, 3–14.

International Tables for Crystallography (2012). Vol. F, Chapter 20.1, pp. 633–641.

PART 20. ENERGY CALCULATIONS AND MOLECULAR DYNAMICS

Chapter 20.1. Molecular-dynamics simulation of protein crystals: convergence of molecular properties of ubiquitin

U. Stocker and W. F. van Gunsteren

20.1.1. Introduction

Molecules in crystals are often believed to have a very rigid structure due to their ordered packing, and the investigation of the molecular motion of such systems is often considered to be of little interest. In contrast to small-molecule crystals, however, the solvent concentration in protein crystals is high, usually with about half of the crystal consisting of water. Thus, in this respect, one can compare protein crystals with very concentrated solutions and expect non-negligible atomic motion. The atomic mobility in proteins can be investigated by experiment (X-ray diffraction, NMR) or by molecular simulation.

Today's experimental techniques are very advanced. They are, however, only able to examine time- and ensemble-averaged structures and properties. In contrast, with simulations one can go beyond averaged properties and examine the motions of a single molecule in the pico- and nanosecond time regime. Such simulations have become possible with the availability of high-resolution structural data, which provide adequate starting structures for biologically relevant systems. Depending on the kind of property in which one is interested, different methods of simulation may be used. Equilibrium properties can be obtained using either Monte Carlo (MC) or molecular-dynamics (MD) simulation techniques, but motions can only be observed with the latter. Current interest in the simulation community mainly focuses on dissolved proteins as they would be in their natural environment. Force fields are parameterized to mimic the behaviour and function of proteins in a solution, and few crystal simulations have been performed. Consequently, a crystal environment provides an excellent opportunity to test a force field on a task for which it should be appropriate, but for which it has not been directly parameterized.

Apart from the analysis of the dynamic properties of a system, MD simulations are also used in structure refinement. In refinement, be it X-ray crystallographic or NMR, a special term is added to the standard physical force field to reflect the presence of experimental data:

$$V(\mathbf{r}) = V^{\rm phys}(\mathbf{r}) + V^{\rm special}(\mathbf{r}). \qquad (20.1.1.1)$$

In NMR, a variety of properties can be measured, and each of these can be used in the definition of an additional term that restrains the generated structures to reproduce given experimental values. Refinement procedures exist that use nuclear-Overhauser-effect (van Gunsteren *et al.*, 1984; Kaptein *et al.*, 1985), *J*-value (Torda *et al.*, 1993) and chemical-shift (Harvey & van Gunsteren, 1993) restraints. In crystallography, X-ray intensities are used to generate the restraining energy contribution (Brünger *et al.*, 1987; Fujinaga *et al.*, 1989). Combined NMR/X-ray refinement uses both solution and crystal data (Schiffer *et al.*, 1994).

As in an experiment, averages over time and molecules are measured, and instantaneous restraints can lead to artificial rigidity in the molecular system (Torda *et al.*, 1990). This can be circumvented by restraining time or ensemble averages, instead of instantaneous values, to the value of the measured quantity. Time averaging has been applied to nuclear Overhauser effects (Torda *et al.*, 1990) and *J* values (Torda *et al.*, 1993) in NMR structure determination and to X-ray intensities in crystallography (Gros *et al.*, 1990; Gros & van Gunsteren, 1993; Schiffer *et al.*, 1995). Ensemble averaging has been applied in NMR refinement (Scheek *et al.*, 1991; Fennen *et al.*, 1995). For a more detailed discussion of restrained MD simulations, we refer to the literature (van Gunsteren *et al.*, 1994, 1997).

The first unrestrained MD simulations of a protein in a crystal were carried out in the early 1980s (van Gunsteren & Karplus, 1981, 1982). The protein concerned was bovine pancreatic trypsin inhibitor (BPTI), a small (58-residue) protein for which high-resolution X-ray diffraction data were available. The initial level of simulation was to neglect solvent, using vacuum boundary conditions. This was improved gradually by the inclusion of Lennard–Jones particles at the density of water as a solvent (van Gunsteren & Karplus, 1982) to let the protein feel random forces and friction from the outside as well as feel a slightly attractive external field. The next step was to use a simple (simple point charge, SPC) water model (van Gunsteren *et al.*, 1983). Further improvement was achieved by incorporating counter ions into the modelled systems to obtain overall charge neutrality (Berendsen *et al.*, 1986).

Despite these early attempts, few unrestrained crystal simulations have been reported in the literature, and, to our knowledge, these involve one to four protein molecules, simulating one unit cell (Shi *et al.*, 1988; Heiner *et al.*, 1992). The maximum time range covered has been less than 100 ps.

In the work described in this chapter, the current state of MD simulation of protein crystals is illustrated. A full unit cell of ubiquitin, containing four ubiquitin and 692 water molecules, has been simulated for a period of two nanoseconds. Since this simulation is an order of magnitude longer than crystal simulations in the literature, it offers the possibility of analysing the convergence of different properties as a function of time and as a function of the number of protein molecules. Converged properties can also be compared with experimental values as a test of the *GROMOS*96 force field (van Gunsteren *et al.*, 1996). Finally, the motions obtained can be analysed to obtain a picture of the molecular behaviour of ubiquitin in a crystalline environment.

20.1.2. Methods

Ubiquitin consists of 76 amino acids with 602 non-hydrogen atoms. Hydrogen atoms attached to aliphatic carbon atoms are incorporated into these (the united-atom approach), and the

Copyright © 2012 International Union of Crystallography

remaining 159 hydrogen atoms are treated explicitly. Ubiquitin crystallizes in the orthorhombic space group $P2_12_12_1$, with a = 5.084, b = 4.277 and c = 2.895 nm. There is one molecule in the asymmetric unit. The protein was crystallized at pH 5.6. The amino acids Glu and Asp were taken to be deprotonated, and Lys, Arg and His residues were protonated, leading to a charge of +1 electron charge per chain. Because this is a small value compared with the size of the system, no counter ions were added. Four chains of ubiquitin, making up a full unit cell of the crystal, were simulated together with 692 water molecules modelled using the SPC water model (Berendsen *et al.*, 1981). 232 water molecules were placed at crystallographically observed water sites, and the remaining 460 were added to obtain the experimental density of 1.35 g cm^{-3}, leading to a system size of 3044 protein atoms and 5120 atoms total.

The crystal structure of ubiquitin [Protein Data Bank (Bernstein *et al.*, 1977) code 1UBQ] solved at 1.8 Å resolution (Vijay-Kumar *et al.*, 1987) was used as a starting point. To achieve the appropriate total density, noncrystallographic water molecules were added, using a minimum distance of 0.220605 nm between non-hydrogen protein atoms or crystallographic water oxygen atoms and the oxygen atoms of the added water molecules, which were taken from an equilibrated water configuration (van Gunsteren *et al.*, 1996). Initial velocities were assigned from a Maxwell–Boltzmann distribution at 300 K. The protein and solvent were coupled separately to temperature baths of 300 K with a coupling time of 0.1 ps (Berendsen *et al.*, 1984). No pressure coupling was applied. Another simulation (results not shown) including pressure coupling showed no significant change in the box volume. Bonds were kept rigid using the *SHAKE* method (Ryckaert *et al.*, 1977), with a relative geometric tolerance of 10^{-4}. Long-range forces were treated using twin range cutoff radii of 0.8 and 1.4 nm (van Gunsteren & Berendsen, 1990). The pair list for non-bonded interactions was updated every 10 fs. No reaction field correction was applied. All simulations were performed using the *GROMOS*96 package and force field (van Gunsteren *et al.*, 1996).

The system was initially minimized for 20 cycles using the steepest-descent method. The protein atoms were harmonically restrained (van Gunsteren *et al.*, 1996) to their initial positions with a force constant of 25000 kJ mol^{-1} nm^{-2}. This minimized structure was then pre-equilibrated in several short MD runs of 500 steps of 0.002 ps each, gradually lowering the restraining force constant from 25000 kJ mol^{-1} nm^{-2} to zero. The time origin was then set to zero, and the entire unit cell was simulated for 2 ns. The time step was 0.002 ps, and every 500th configuration was stored for evaluation. The first 400 ps of the run were treated as equilibration time, the remaining 1.6 ns were used for analysis.

20.1.3. Results

20.1.3.1. Energetic properties

In Fig. 20.1.3.1, the non-bonded contributions to the total potential energy are shown. The non-bonded interactions comprise Lennard–Jones and electrostatic interactions. Solvent–solvent, solute–solute and solute–solvent interaction energies are shown separately. All of these appear converged after approximately 100 ps. The solvent–solvent energy remains close to its initial value during the whole simulation, the water molecules having relaxed during the pre-equilibration period, while the protein was restrained. The protein internal energy increases during the first few hundred picoseconds, but this is compensated by a decrease in the protein–solvent energy as the protein adapts to the force field and the pre-relaxed solvent environment. This effect becomes negligible after about 200 ps, from which time point the system can be viewed as equilibrated with respect to the energies. The distribution of kinetic *versus* potential energy and the total (bonded and non-bonded) energy of the system relaxes even faster (results not shown).

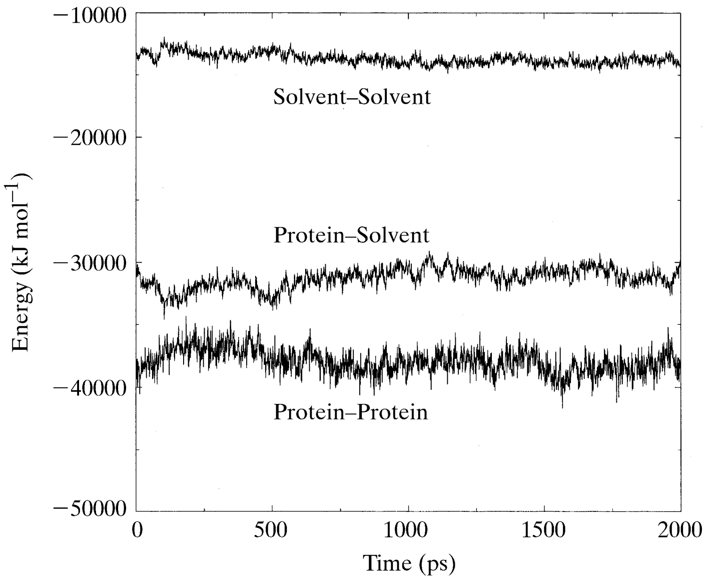


Figure 20.1.3.1
Non-bonded energies (in kJ mol^{-1}) of the simulated system as a function of time.

20.1.3.2. Structural properties

Not all properties converge as fast as the energies. Fig. 20.1.3.2 shows the root-mean-square atom-position deviation (RMSD) from the X-ray structure for each of the four individual chains for both Cα atoms and all atoms. Here, several relaxation periods can be distinguished. After the initial increase, which occurs during the first 50 ps of the simulation, a plateau is reached, and the

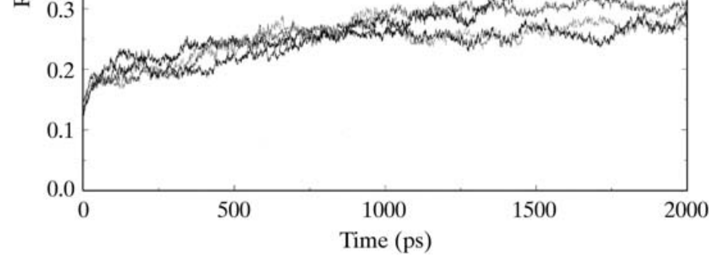


Figure 20.1.3.2
Root-mean-square atom-positional deviations (RMSD) in nm from the X-ray structure of the four different protein molecules in the unit cell as a function of time. Rotational and translational fitting was applied using the Cα atoms of residues 1–72. The upper and lower graphs show the deviations for the Cα atoms and for all atoms, respectively.

system is apparently stable until 300 ps. The values reached are 0.12 nm for the Cα atoms and 0.20 nm if all atoms are considered. These numbers are comparable with results obtained in crystal simulations of other proteins of equivalent length reported in the literature (van Gunsteren *et al.*, 1983; Berendsen *et al.*, 1986; Shi *et al.*, 1988; Heiner *et al.*, 1992; Levitt *et al.*, 1995). After 300 ps, however, the values increase slowly again. For the Cα atoms, there is apparently a second plateau from 300 to 850 ps, but during this period the RMSD for all atoms continues to increase monotonically. After 850 ps, a final plateau is reached. During the second nanosecond of the simulation (1000–2000 ps), the RMSDs are 0.21 nm for the Cα atoms and 0.29 nm for all atoms. The RMSD of chain 1 is an exception. There is a strong increase after 1200 ps due to a movement of a particular part of the chain which will be addressed later. To ensure that the RMSD values have converged, longer runs would be required.

Although the RMSD values shown in Fig. 20.1.3.2 grow larger than those usually observed in the course of short simulations, the hydrogen-bonding pattern and thus the secondary-structure elements observed in the X-ray structure are reproduced well (Table 20.1.3.1). Most of the hydrogen bonds reported (Vijay-Kumar *et al.*, 1987) show high occupancies during the whole simulation, especially the ones within secondary-structure elements. Only six out of the 44 hydrogen bonds present in the X-ray structure are disrupted during the simulation. Hydrogen bonds in the α-helix (residues 23–34) show very high occupancies, ranging from 75% at the N-terminus to well over 90% inside the helix. Only its C-terminus shows signs of instability, with the α-helix deforming towards a 3_{10}-helix. The β-sheet pattern is, apart from chain 1 in the region of residues 49–64, as stable as the α-helix. Occupancies range from 55% up to 95%. The six hydrogen bonds not reproduced (five 3_{10} and one α-helical) can be rationalized as follows. The bridges 10–7 and 65–62 are part of the most mobile regions of the protein. These regions involve residues 7–10, 51–54 and 62–65 (Vijay-Kumar *et al.*, 1987). The hydrogen bond at the C-terminal end of the α-helix (35–31) is lost, and the preceding hydrogen bond (34–30) is partly changed, indicating that the end of the α-helix is deformed towards a 3_{10}-helix. The donor of the bond 40–37 is replaced by residue 41, and the four-residue 3_{10}-helix that was stabilized by hydrogen bonds 58–55 and 59–56 is replaced by an α-helical hydrogen bond 59–55. The high occupancy of this particular bond and the complete absence of the two observed experimentally indicate an early rearrangement in this part of the structure (before the analysis period) which is stable during the rest of the simulation.

Backbone–side-chain hydrogen bonds are less well reproduced than backbone–backbone interactions. While some are present 80–90% of the time, others are present less than 50% of the time. Two out of the seven hydrogen bonds in which a backbone atom is the donor are not observed in the simulation; both involve the OG1 atom of Thr7 as an acceptor. The hydrogen-donor atoms are the backbone nitrogen atoms of residues Thr9 and Lys11, both of which have high experimental *B* factors (18.32 Å^2 for Thr9 and 13.56 Å^2 for Lys11). The mean of the experimental *B* factors is 10.73 Å^2 for the backbone atoms and 13.41 Å^2 for all protein atoms. Where a side-chain atom is the donor, three out of the five hydrogen bonds present in the X-ray structure are not found in the simulation. All of these involve the side-chain nitrogen atom of a lysine residue as the donor, the experimental *B* factors of which range from 23.92 Å^2 for the NZ atom of Lys48 up to 30.06 Å^2 for the NZ atom of Lys33. Of the four side-chain–side-chain hydrogen bonds, not one is observed as in the crystal. The 54–58 hydrogen bond is, however, replaced by a 55–58 hydrogen bond. All the others involve very mobile atoms with large experimental *B* factors as donors and acceptors.

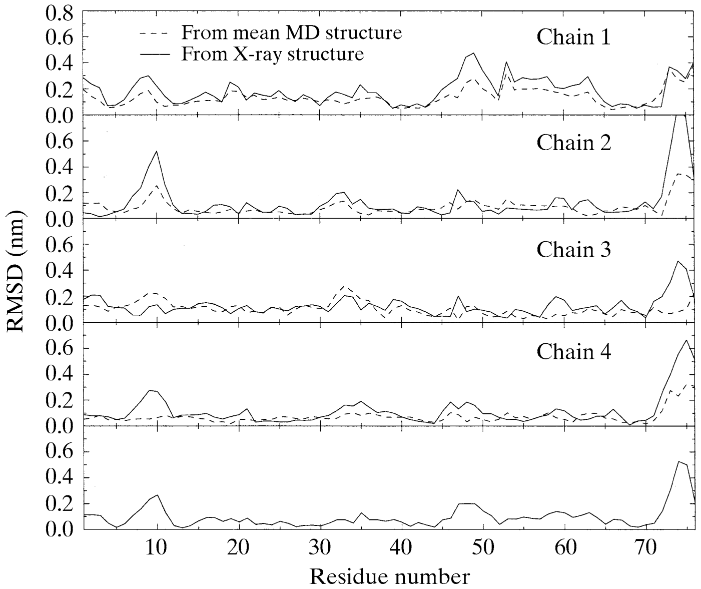


Figure 20.1.3.3
Root-mean-square Cα-atom-position deviation (RMSD) in nm from a reference structure as a function of the residue number using the final 1.6 ns of the simulation. RMSDs of the four protein molecules in the unit cell from the mean molecular-dynamics structure (dashed line) and from the X-ray structure (solid line) are shown in the first four graphs. The bottom graph shows the RMSD of the mean (over the four molecules) MD structure from the X-ray structure.

There is one intermolecular hydrogen bond (Table 20.1.3.2) in the starting structure which is not seen in the simulation. The donor is the side-chain nitrogen atom of Lys6, which has an experimental *B* factor of 20.55 Å^2, and the acceptors are the side-chain oxygen atoms of Glu51, with *B* factors of 32.13 and 33.44 Å^2. Most of the hydrogen bonds not reproduced in the simulation contain at least one mobile atom. Although these atoms do not remain fixed at their equilibrium positions, they may still stabilize the structure on average.

In Fig. 20.1.3.3, the deviation of the Cα atoms of the different chains from the X-ray structure and from the mean MD structure are presented together with the deviation of the mean MD structure from the X-ray structure. Overall, the individual protein molecules remain close to the experimental structure; however, parts of the structure do deviate substantially. The region involving residues 7–11, which experimentally has high *B* factors (implying high mobility), has, in three out of the four cases, an RMSD for Cα atoms of 0.3 nm or greater. Chain 3, in contrast, is close to the X-ray structure, and the mean MD structure is closer to the X-ray structure than are any of the individual chains. This suggests that the simulation does not deviate systematically from the X-ray structure, but rather that different regions of conformational space are sampled by the different molecules. The same holds for the very mobile region between residues 47 and 64, where chain 1 deviates dramatically from both the mean MD structure and the X-ray structure. In the other chains, this part of the protein remains close to the X-ray structure. Overall, the deviation from the X-ray structure is largest in the C-terminal region. This part is also ill-defined in the experiment, with occupancies of 0.45 for residues 73 and 74, and 0.25 for the terminal two glycines. Other parts of the protein, especially the stable secondary-structure elements, stay close to the X-ray structure. In the average structure, the α-helix, including its C-terminal part which was deformed to a 3_{10}-helix, deviates by a

Table 20.1.3.1

Occurrence of intramolecular hydrogen bonds (%) during the final 1.6 ns of the simulation

The criteria for a hydrogen bond to be present are angle donor–hydrogen–acceptor ≥ 135°, distance hydrogen–acceptor ≤ 0.25 nm. Hydrogen bonds are shown if they are either present in the X-ray structure or if at least one of the four protein molecules in the unit cell shows the hydrogen bond of interest for at least 50% of the simulation time. The letter h appended to an amino-acid code indicates that the residue is protonated.

Hydrogen bonds			Molecular dynamics			
Backbone	Backbone	X-ray structure	Molecule 1	Molecule 2	Molecule 3	Molecule 4
3Ile H	15Leu O	100	94	94	95	98
4Phe H	65Ser O	100	85	69	87	77
5Val H	13Ile O	100	80	90	87	93
6Lysh H	67Leu O	100	85	82	88	94
7Thr H	11Lysh O	100	65	49	54	62
8Leu H	69Leu O	0	5	52	19	55
10Gly H	7Thr O	100	0	0	0	0
13Ile H	5Val O	100	86	76	70	87
15Leu H	3Ile O	100	87	92	72	82
17Val H	1Met O	100	68	39	79	51
21Asp H	18Glu O	100	68	84	84	90
23Ile H	54Arg O	100	0	74	89	92
24Glu H	52Asp O	100	58	69	63	84
26Val H	22Thr O	100	92	69	78	61
27Lysh H	23Ile O	100	94	97	98	99
28Ala H	24Glu O	100	71	71	84	89
29Lysh H	25Asn O	100	91	79	94	88
30Ile H	26Val O	100	92	76	94	91
31Gln H	27Lysh O	100	85	53	66	93
32Asp H	28Ala O	100	82	27	87	77
33Lysh H	29Lysh O	100	23	13	81	51
33Lysh H	30Ile O	0	59	23	7	19
34Glu H	30Ile O	100	95	54	64	86
35Gly H	31Gln O	100	0	0	0	0
36Ile H	34Glu O	0	62	50	28	35
40Gln H	37Pro O	100	0	0	0	0
41Gln H	37Pro O	0	68	56	72	20
41Gln H	38Pro O	100	14	25	14	50
42Arg H	70Val O	100	82	82	83	88
44Ile H	68Hish O	100	84	96	93	95
45Phe H	48Lysh O	100	20	74	77	91
48Lysh H	45Phe O	100	24	62	59	44
50Leu H	43Leu O	100	29	88	92	85
54Arg H	51GluO	100	20	60	19	69
56Leu H	21Asp O	100	0	90	81	81
57Ser H	19Pro O	100	3	78	86	83
58Asp H	55Thr O	100	0	0	0	0
59Tyr H	55Thr O	100	58	86	92	85
59Tyr H	56Leu O	100	0	0	0	0
60Asn H	57Ser O	100	38	34	60	58
61Ile H	56Leu O	100	67	7	63	56
64Glu H	2Gln O	100	0	42	6	95
65Ser H	62Gln O	100	0	0	0	0
67Leu H	4Phe O	100	69	74	87	70
68Hish H	44Ile O	100	62	68	83	89
69Leu H	6Lysh O	100	79	72	92	90
70Val H	42Arg O	100	91	89	90	91
72Arg H	40Gln O	100	79	59	85	78

maximum of 0.08 nm from the X-ray structure, although, as seen before, the individual chains may deviate more. The β-sheet regions also stay close to the X-ray structure. As with the helix, residues 1–7, 40–45 and 64–72 stay within 0.1 nm RMSD from the X-ray structure. The β-strands formed by residues 10–17 and 48–50 are not as similar to the experiment, since they lie close to mobile regions and are thus influenced by neighbouring mobile residues. For the strand formed by residues 10–17, from residue 12 onwards the same structural similarity is reached as for all other secondary-structure elements, and residues 48–50 are, again, strongly influenced by the moving part of chain 1.

20.1.3.3. Effect of the translational and rotational fitting procedure

In Fig. 20.1.3.4, the impact of different fitting protocols on atomic mean-square position fluctuations (RMSFs) is examined. B factors are related to mean-square position fluctuations according to

$$B_i = (8\pi^2/3)\big\langle(\mathbf{r}_i - \langle\mathbf{r}_i\rangle)^2\big\rangle, \qquad (20.1.3.1)$$

where the angle brackets indicate a time or a combined time and ensemble average. Molecule 4 was selected because it is the most stable. RMSFs of the Cα atoms were calculated directly from the simulation trajectory (Fig. 20.1.3.4*a*) after applying a translational fit using the Cα atoms of residues 1–72, which are well defined in the X-ray structure (Fig. 20.1.3.4*b*), and after applying both a rotational and a translational fit on residues 1–72 (Fig. 20.1.3.4*c*). In Fig. 20.1.3.4(*d*), a translational and rotational fit was applied to all Cα atoms (residues 1–76). Removal of the overall translational component of motion reduces the positional fluctuations by 0.04 nm on average. Only the RMSFs in the proximity of the

Table 20.1.3.1 (continued)

Hydrogen bond		X-ray structure	Molecular dynamics			
Backbone	Side chain		Molecule 1	Molecule 2	Molecule 3	Molecule 4
2Gln H	64Glu OE2	0	63	7	84	0
9Thr H	7Thr OG1	100	0	0	0	0
11Lysh H	7Thr OG1	100	0	0	0	0
18Glu H	21Asp OD2	100	80	3	0	0
20Ser H	18Glu OE2	0	0	0	55	0
25Asn H	22Thr OG1	100	31	13	61	38
51Glu H	59Tyr OH	100	46	87	56	76
55Thr H	58Asp OD1	100	29	62	22	75
58Asp H	55Thr OG1	100	53	76	72	86
64Glu H	64Glu OE2	0	55	6	16	0

Hydrogen bond		X-ray structure	Molecular dynamics			
Side chain	Backbone		Molecule 1	Molecule 2	Molecule 3	Molecule 4
29Lysh HZ2	16Glu O	100	0	0	0	0
33Lysh HZ2	14Thr O	100	0	0	0	0
41Gln HE21	27Lysh O	100	81	91	47	71
41Gln HE22	36Ile O	100	90	89	60	83
48Lysh HZ3	46Ala O	100	0	0	0	0

Hydrogen bond		X-ray structure	Molecular dynamics			
Side chain	Side chain		Molecule 1	Molecule 2	Molecule 3	Molecule 4
11Lysh HZ2	34Glu OE2	100	0	0	0	0
20Ser HG	18Glu OE2	0	0	0	60	0
27Lysh HZ2	52Asp OD2	100	0	0	0	0
49Gln HE21	16Glu OE1	100	0	0	0	0
54Arg HH12	58Asp OD1	100	0	0	0	0
55Thr HG1	58Asp OD1	0	44	83	29	86

end of the large α-helix formed by residues 23–34 are not affected by the introduction of translational fitting. In contrast, it is exactly this region where the fluctuations are substantially lowered by introducing an additional rotational fit. The regions before residue 27 and after residue 42 are only slightly affected by the removal of overall rotation. These findings suggest that the entire protein translates by about 0.04 nm, while the α-helix region remains close to its initial position, thus rotating relative to the rest of the protein. Inclusion of the four C-terminal residues in the fitting procedure only affects the RMSFs of these residues and residues in the rotating part of the molecule, indicating that these four residues move together with the rest of the molecule. The atom-position fluctuations obtained by applying a full (rotational and translational) fit are determined by internal motion only. The largest RMSFs for residues 1–72 are 0.12 nm. RMSFs of the two C-terminal glycines are 0.26 nm if the last four residues are excluded from the fitting and 0.22 nm otherwise.

If the same properties are examined but averaged over all the chains, similar trends can be observed (Fig. 20.1.3.5). If no fitting is applied, the RMSFs of 0.24 nm, on average, indicate that the different molecules show relative translation and rotation. After translational fitting is applied, the mean RMSFs drop to 0.18 nm. Thus, the molecules translate within the unit cell. If, in addition, the rotational component of overall motion is removed, the whole helix region is much less mobile than before, and the mean RMSFs drop to 0.14 nm. The same holds for region 47–64, dominated by the rotation of part of chain 1. Fluctuations are generally much larger than before when only chain 4 was observed, again indicating that the distinct chains behave in an uncorrelated way. The size of the peaks of the RMSFs averaged over all chains is around 0.22 nm, compared with 0.27 nm when overall rotation is still present. Thus, in addition to internal rotations, relative rotations of the different molecules occur. If the fit is not applied only to the well defined Cα atoms of residues 1–72, the RMSF value becomes slightly higher – apart from the C-terminal region – but this effect is small, with the mean RMSF staying at 0.14 nm. However, the relative heights of the peaks differ, which shows that it is crucial to define a standard fitting

Table 20.1.3.2
Occurrence of intermolecular hydrogen bonds (%) during the final 1.6 ns of the simulation

The criteria for a hydrogen bond to be present are: angle donor–hydrogen–acceptor $\geq 135^\circ$, distance hydrogen–acceptor ≤ 0.25 nm. Hydrogen bonds are shown if they are either present in the X-ray structure or if at least one of the four protein molecules in the unit cell shows the hydrogen bond of interest for at least 50% of the simulation time.

Hydrogen bond		X-ray structure	Molecular dynamics			
			Molecules 1–4	Molecules 2–3	Molecules 3–2	Molecules 4–1
6 Lys HZ3	51 Glu OE1	100	0	0	0	0
12 Thr HG1	18 Glu OE1	0	56	57	75	34
49 Gln H	8 Leu O	0	10	34	0	67
68 Hish HE2	32 Asp OD2	0	0	0	53 (3–1)	13 (4–2)
71 Leu H	58 Asp O	0	65	0	0	0

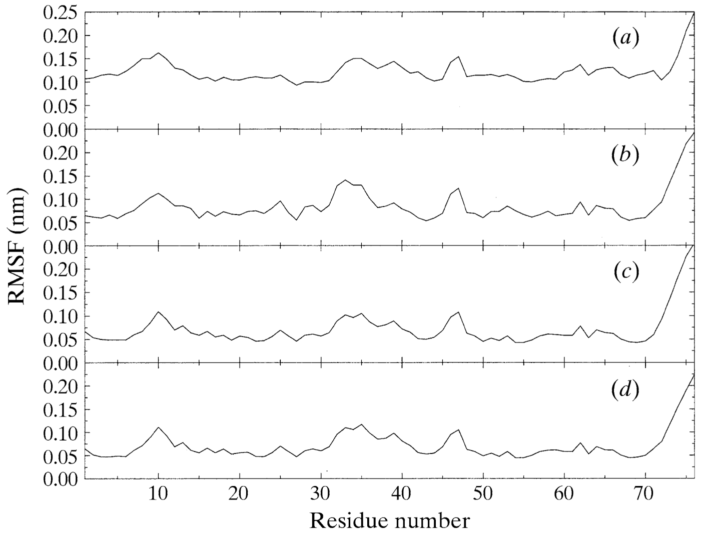


Figure 20.1.3.4
Root-mean-square Cα-atom-position fluctuations (RMSFs) in nm are shown for molecule 4 as a function of residue number. RMSFs are averaged over the final 1.6 ns. In (*a*), no fitting was applied; in (*b*), translational fitting was applied using the Cα atoms of residues 1–72; and in (*c*), the rotational component of motion was also removed. In (*d*), translational and rotational fitting was applied over all Cα atoms (1–76).

protocol that must not be changed during the course of the analysis.

20.1.3.4. Effect of the averaging period

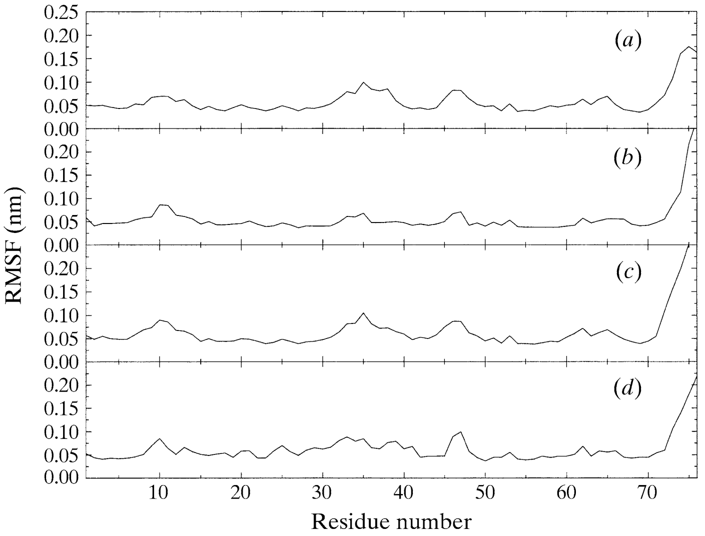


Figure 20.1.3.6
Root-mean-square Cα-atom-position fluctuations (RMSFs) in nm are shown for molecule 4, with full translational and rotational fitting over the Cα atoms of residues 1–72. Different averaging periods are compared. (*a*) shows the RMSFs for the period 400–800 ps, (*b*) shows those for the period 800–1200 ps. In (*c*), the results for the previous two periods are averaged (400–1200 ps), and in (*d*), the results for the period 1200–2000 ps are shown.

In Fig. 20.1.3.6, we concentrate again on molecule 4. Comparing different averaging periods of 400 and 800 ps with different starting points, it can be seen that, in general, the later the simulation, the less motion observed. During the period 800–1200 ps, only the small region between residues 9 and 12 shows more mobility than that between 400 and 800 ps. In the stable region, the rest of the molecule shows the same mobility as in the earlier time period. In the parts that are most mobile between 400 and 800 ps, the motions decrease significantly after the latter time point, indicating that equilibrium is reached. Focusing on the longer averaging periods, 400–1200 ps *versus* 1200–2000 ps, we see that over the whole chain mobility is comparable, indicating clear equilibrium as far as internal motions are concerned. The fluctuations during the 400–1200 ps period are of the same size as those of the shorter subperiods, 400–800 ps and 800–1200 ps. They are thus determined by movements on a timescale shorter than 400 ps.

In Fig. 20.1.3.7, the effect of different averaging periods is again considered, but taking the whole unit cell into account. Comparing RMSFs between 400 and 800 ps with those of the following 400 ps period, no significant difference is seen. In fact, contrary to what was observed when chain 4 was examined, somewhat more mobility is evident in the later period compared with the earlier one. This difference shows that although the configurations of the single chains converge rapidly, the different chains visit different parts of phase space. The fact that the fluctuations in the period 400–1200 ps are between those of the two shorter analysis periods is another indication that the single-

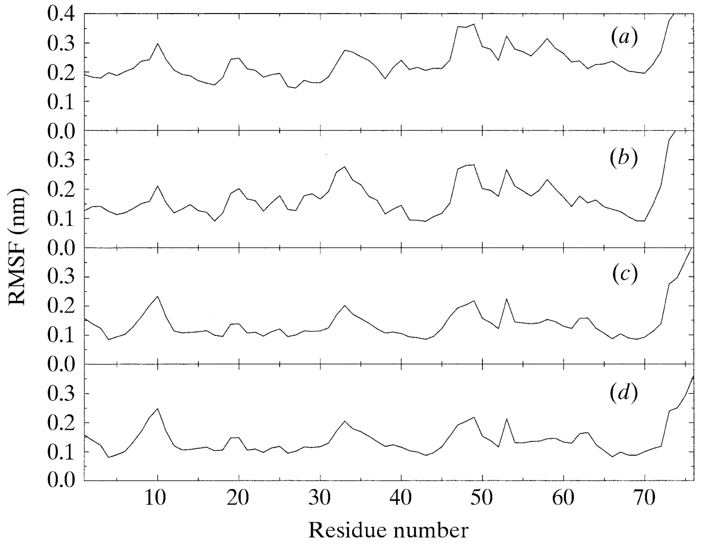


Figure 20.1.3.5
Root-mean-square Cα-atom-position fluctuations (RMSFs) in nm are shown using the same fitting protocols as in Fig. 20.1.3.4, but averaged over all four protein molecules in the unit cell.

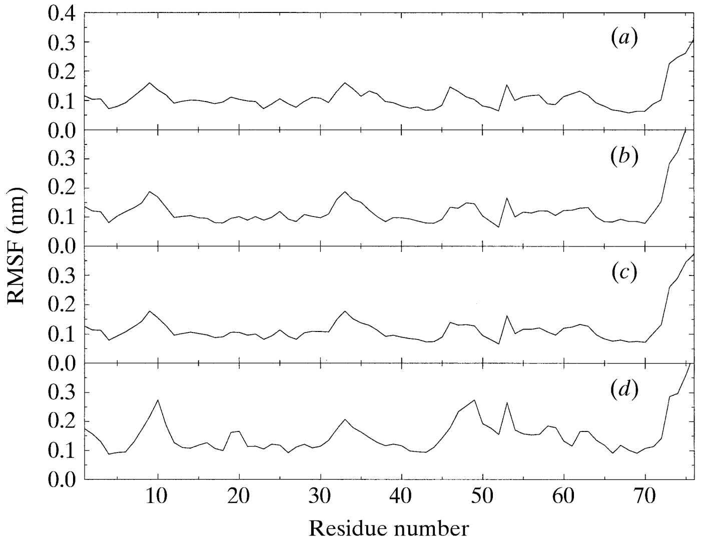


Figure 20.1.3.7
Root-mean-square Cα-atom-position fluctuations (RMSFs) in nm are shown using the same averaging periods as in Fig. 20.1.3.6, but averaged over all four protein molecules in the unit cell.

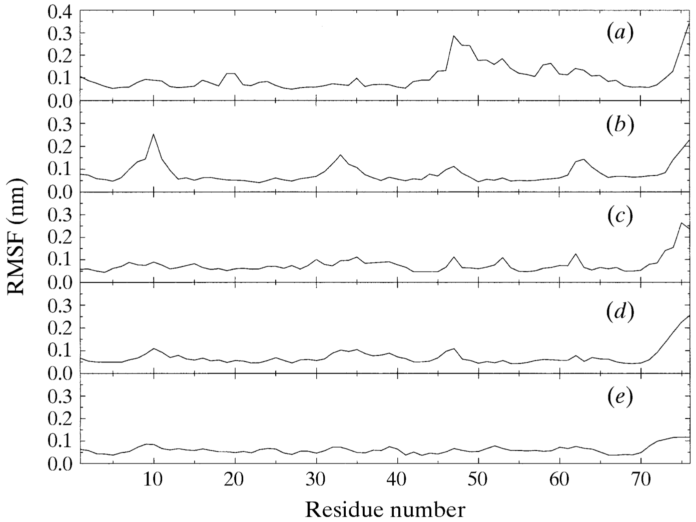


Figure 20.1.3.8

Root-mean-square Cα-atom-position fluctuations (RMSFs) in nm for the four protein molecules in the unit cell as a function of the residue number. Full translational and rotational fitting was applied to the Cα atoms of residues 1–72 using the final 1.6 ns of the simulation [(*a*)–(*d*)]. (*e*) shows the corresponding values defined by equation (20.1.3.1), obtained from experimental *B* factors.

chain movements are converged on these short timescales. In the last 800 ps of the simulation, the RMSFs are substantially higher than in the 800 ps window before that. All of the peaks can be traced back to one of the single chains. If only one of the four molecules differs strongly from the other three, this one determines the magnitude of the fluctuations of the average. The peak at residue 10 comes from chain 2, the ones around residue 20 and the whole region 47–64 are determined by chain 1. The peak at residue 33 originates in chain 3, which at this point differs substantially from the mean MD structure (Fig. 20.1.3.3).

20.1.3.5. Internal motions of the proteins

Fig. 20.1.3.8 displays the atomic root-mean-square position fluctuations for the Cα atoms of the four protein molecules during the whole analysis period, together with corresponding values obtained using equation (20.1.3.1) and the crystallographic *B* factors. Rotational and translational fitting was applied using the Cα atoms of residues 1–72, and the fluctuations were averaged over the final 1.6 ns. The mobility of the stable secondary-structure elements in the simulation is comparable with that inferred from the experiment. There is a correlation between the more mobile parts of the proteins in the simulation and large *B* factors in the X-ray structure, but the magnitude of the fluctuations is overestimated in the simulation. The movements of the single chains can be rationalized as follows. In chain 1, the whole region from Gly47 onwards rotates around a stable axis formed by residues 41–46. This part lies, as do all the flexible regions, on the exterior of the protein. Residues 19 and 20, which are stable in all but this single chain, are in contact with this moving part. This rotation, which tends to compact the protein, occurs during the 200 ps period between 1350 and 1550 ps after the start of the simulation, in which the atom-position RMSD from the X-ray structure increases significantly (Fig. 20.1.3.2). Overall, chain 2 is more stable than chain 1. Nevertheless, the end of the unwinding helix shows large fluctuations. In the course of this deformation, the side-chain nitrogen atom of Lys11 moves from close to the OE atom of Glu34 towards the backbone oxygen atom of Lys33, which is associated with a change in the position of Gly10. A similar but smaller motion occurs in chain 4. Both lysines, Lys33 and Lys63, are fully exposed to the solvent and have no intramolecular contacts. In chain 3, the flexible residues are also not part of secondary-structure elements and are located on the outside of the protein. The backbone oxygen atom of Gln62 that moves in all the four chains has, in addition, the closest contact to another heavy atom: the OG1 atom of Ser65 is only 2.51 Å away, and the van der Waals repulsion of these atoms causes them to move further away from one another. The mobile residues in chain 4 are again in contact with the solvent, Gly35, Gly47, Gln62, the end of the helix and Gly10. The terminal residues of all the protein molecules are very mobile, as observed experimentally in the crystal.

Table 20.1.3.3

Root-mean-square fluctuations of polypeptide backbone and ψ dihedral angles (°) for the different molecules using different time-averaging periods

The bottom row shows the averages over all four protein molecules.

Molecule	400–800 ps φ/ψ	400–1200 ps φ/ψ	400–2000 ps φ/ψ
1	18.4/22.9	19.5/23.7	31.0/33.6
2	17.2/17.0	18.6/18.7	23.6/26.8
3	18.5/20.4	25.6/26.3	35.3/37.5
4	19.7/18.8	19.4/20.3	21.6/28.8
All	26.1/26.2	28.0/28.6	35.2/38.1

20.1.3.6. Dihedral-angle fluctuations and transitions

Backbone dihedral-angle fluctuations and transitions are examined in Tables 20.1.3.3 and 20.1.3.4 using different analysis periods. After the first 400 ps of analysis, the φ/ψ dihedral-angle fluctuations differ only slightly, but if longer averaging times are chosen, the different protein molecules show larger differences

Table 20.1.3.4

Number of protein-backbone dihedral-angle transitions per 100 ps for the different molecules using different time periods

Dihedral angles with threefold and sixfold potential-energy wells are distinguished. The bottom rows show the averages over all protein molecules.

(*a*) 120° transitions

Molecule	400–800 ps	400–1200 ps	400–2000 ps
1	46.5	45.4	47.7
2	40.5	41.5	47.3
3	50.5	57.1	51.3
4	44.8	46.4	46.4
All	45.6	47.6	48.2

(*b*) 60° transitions

Molecule	400–800 ps	400–1200 ps	400–2000 ps
1	245.5	246.6	289.3
2	271.5	272.1	261.3
3	381.5	381.0	348.3
4	356.8	325.4	325.4
All	313.8	306.3	306.1

(*c*) All transitions

Molecule	400–800 ps	400–1200 ps	400–2000 ps
1	292.0	292.0	336.9
2	312.0	313.7	308.6
3	432.0	438.1	399.6
4	401.5	371.8	371.8
All	359.4	353.9	354.2

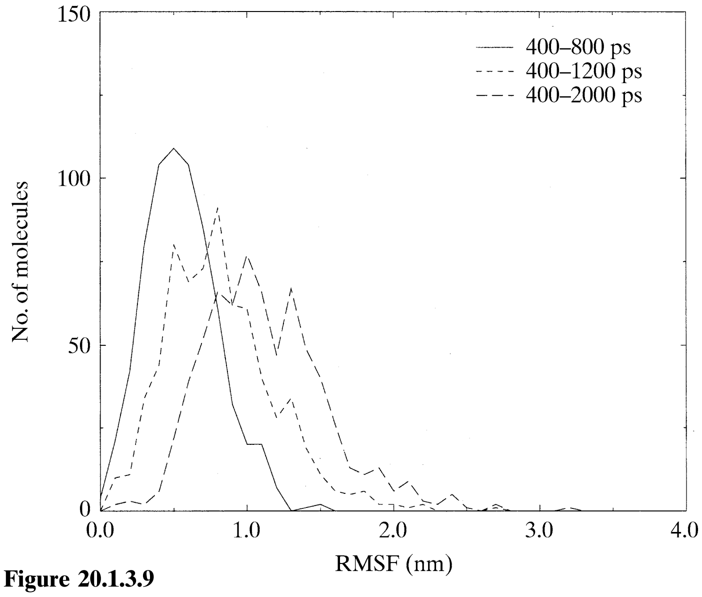


Figure 20.1.3.9
The number of water molecules with a given root-mean-square oxygen-position fluctuation (RMSF) in nm are shown for different averaging periods: 400–800 ps (solid line), 400–1200 ps (short-dashed line), 400–2000 ps (long-dashed line).

from one another. These fluctuations also increase for longer analysis times, indicating that they are not yet converged after 2 ns. In the period from 800 to 1200 ps, chain 3 shows a large increase in mean-square dihedral-angle fluctuations, whereas the Cα-atom-position RMSDs with respect to the X-ray structure during the same time fluctuate around a plateau. Thus, there is a lot of flexibility without the simulation structure diverging from the experimental one. Protein molecule 3, for example, shows the largest φ/ψ fluctuations of all the four molecules, and it shows the lowest atom-position RMSDs of Cα atoms from the X-ray structure at the end of the simulation (Fig. 20.1.3.2), indicating that it explores phase space around the equilibrium structure. If, in contrast, the Cα-atom-position RMSDs, with respect to the X-ray structure, increase significantly, larger dihedral-angle fluctuations are also observed, for example, in molecule 1 after 1200 ps.

Concerning relaxation, observations similar to those made before can be made when analysing dihedral-angle transitions (Table 20.1.3.4). The number of transitions for the different chains differs by about 20%. Within a single chain, however, the number of transitions increases in proportion to the observation time. Again, the protein molecules showing the most transitions do not have the largest Cα-atom-position RMSDs from the X-ray structure. Thus, only certain dihedral-angle flips induce large movements that determine the RMSD value.

20.1.3.7. Water diffusion

In Fig. 20.1.3.9, the number of water oxygen atoms with a given atomic root-mean-square position fluctuation (RMSF) are plotted. The time evolution and the shapes of these curves are similar to those obtained for bulk water, a Gaussian distribution with the maximum at larger RMSF values and larger standard deviations when using longer averaging times. Using a diffusion constant of bulk SPC water at 300 K of 3.9×10^{-3} nm^2 ps^{-1} (Smith & van Gunsteren, 1995), the root-mean-square position fluctuation for an average water molecule would be 1.25 nm for a 400 ps period, 1.77 nm for a 800 ps period, and 2.5 nm for a 1600 ps period. Comparison of these values with the distributions in Fig. 20.1.3.9 indicates that the motion of most of the crystal water molecules is restricted by the crystalline environment. A number of water molecules adopt well defined positions for periods of the order of some 100 ps. Several water molecules were also observed to visit the same location repeatedly. This indicates that, although not remaining fixed in space, they stay close to a crystallographically determined site which they revisit periodically, alternating with other water molecules.

20.1.4. Conclusions

In the present molecular-dynamics simulation, fast convergence in energy, within about 100 ps, was observed. Other properties, such as dihedral-angle fluctuations and backbone atom-position fluctuations, converged on an intermediate timescale of hundreds of picoseconds. Root-mean-square deviations of the simulated protein molecules from the starting X-ray structure required of the order of 1 ns to reach a plateau. Longer simulations would be necessary to obtain convergence for all molecular properties. The convergence of quickly relaxing properties of the system, such as the energies, is not a reliable indicator of the degree of global convergence in such a molecular-dynamics simulation.

The *GROMOS*96 force field used in this simulation largely reproduces the secondary structure and the relative internal mobility of ubiquitin. The simulation does, however, overestimate the magnitude of the fluctuations in the most mobile regions of the protein. The different protein molecules were observed to translate and rotate relative to one another during this simulation. This indicates that the force field would not be able to reproduce the experimental melting temperature of this crystal under the conditions simulated.

The authors wish to thank Dr Thomas Huber for fruitful discussions and Dr Alan Mark for critical reading of the manuscript. Financial support was obtained from the Schweizerischer Nationalfonds (Project 21-41875.94), which is gratefully acknowledged.

References

Berendsen, H. J. C., van Gunsteren, W. F., Zwinderman, H. R. J. & Geurtsen, R. (1986). *Simulations of proteins in water. Ann. N. Y. Acad. Sci.* **482**, 269–285.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). *Molecular dynamics with coupling to an external bath. J. Chem. Phys.* **81**, 3684–3690.

Berendsen, H. J. C., Postma, J. P. M.,van Gunsteren, W. F. & Hermans, J. (1981). *Interaction models for water in relation to protein hydration*. In *Intermolecular Forces*, edited by B. Pullman, pp. 331–342. Dordrecht: Reidel.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R-factor refinement by molecular dynamics. Science*, **235**, 458–460.

Fennen, J., Torda, A. E. & van Gunsteren, W. F. (1995). *Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. J. Biomol. NMR*, **6**, 163–170.

Fujinaga, M., Gros, P. & van Gunsteren, W. F. (1989). *Testing the method of crystallographic refinement using molecular dynamics. J. Appl. Cryst.* **22**, 1–8.

Gros, P. & van Gunsteren, W. F. (1993). *Crystallographic refinement and structure-factor time-averaging by molecular dynamics in the absence of a physical force field. Mol. Simul.* **10**, 377–395.

Gros, P., van Gunsteren, W. F. & Hol, W. G. J. (1990). *Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science*, **249**, 1149–1152.

Gunsteren, W. F. van & Berendsen, H. J. C. (1990). *Computer simulations of molecular dynamics: methodology, applications and perspectives in chemistry*. *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023.

Gunsteren, W. F. van, Berendsen, H. J. C., Hermans, J., Hol, W. G. J. & Postma, J. P. M. (1983). *Computer simulation of the dynamics of hydrated protein crystals and its comparison with X-ray data*. *Proc. Natl Acad. Sci. USA*, **80**, 4315–4319.

Gunsteren, W. F. van, Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P. & Tironi, I. G. (1996). *Biomolecular Simulation: the GROMOS96 Manual and User Guide.* Vdf Hochschulverlag, Zürich, Switzerland.

Gunsteren, W. F. van, Bonvin, A. M. J. J., Daura, X. & Smith, L. J. (1997). *Aspects of modelling biomolecular structures on the basis of spectroscopic or diffraction data*. In *Modern Techniques in Protein NMR*, edited by Krishna & Berliner. Plenum.

Gunsteren, W. F. van, Brunne, R. M., Gros, P., van Schaik, R. C., Schiffer, C. A. & Torda, A. E. (1994). *Accounting for molecular mobility in structure determination based on nuclear magnetic resonance spectroscopic and X-ray diffraction data*. *Methods Enzymol.* **239**, 619–654.

Gunsteren, W. F. van, Kaptein, R. & Zuiderweg, E. R. P. (1984). *Use of molecular dynamics computer simulations when determining protein structure by 2D-NMR*. In *Proceedings of the NATO/CECAM Workshop on Nucleic Acid Conformation and Dynamics*, edited by W. K. Olson, pp. 79–97. France: CECAM.

Gunsteren, W. F. van & Karplus, M. (1981). *Effect of constraints, solvent and crystal environment on protein dynamics*. *Nature (London)*, **293**, 677–678.

Gunsteren, W. F. van & Karplus, M. (1982). *Protein dynamics in solution and in crystalline environment: a molecular dynamics study*. *Biochemistry*, **21**, 2259–2274.

Harvey, T. S. & van Gunsteren, W. F. (1993). *The application of chemical shift calculation to protein structure determination by NMR*. In *Techniques in Protein Chemistry IV*, pp. 615–622. New York: Academic Press.

Heiner, A. P., Berendsen, H. J. C. & van Gunsteren, W. F. (1992). *MD simulation of subtilisin BPN′ in a crystal environment*. *Proteins*, **14**, 451–464.

Kaptein, R., Zuiderweg, E. R. P., Scheek, R. M., Boelens, R. & van Gunsteren, W. F. (1985). *A protein structure from nuclear magnetic resonance data, lac repressor headpiece*. *J. Mol. Biol.* **182**, 179–182.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). *Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution*. *Comput. Phys. Commun.* **91**, 215–231.

Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. (1977). *Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes*. *J. Comput. Phys.* **23**, 327–341.

Scheek, R. M., Torda, A. E., Kemmink, J. & van Gunsteren, W. F. (1991). *Structure determination by NMR: the modelling of NMR parameters as ensemble averages*. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy*, edited by J. C. Hoch, F. M. Poulsen & C. Redfield, NATO ASI Series, Vol. A225, pp. 209–217. New York: Plenum Press.

Schiffer, C. A., Gros, P. & van Gunsteren, W. F. (1995). *Time-averaging crystallographic refinement: possibilities and limitations using α-cyclodextrin as a test system*. *Acta Cryst.* D**51**, 85–92.

Schiffer, C. A., Huber, R., Wüthrich, K. & van Gunsteren, W. F. (1994). *Simultaneous refinement of the structure of BPTI against NMR data measured in solution and X-ray diffraction data measured in single crystals*. *J. Mol. Biol.* **241**, 588–599.

Shi, Y.-Y., Yun, R.-H. & van Gunsteren, W. F. (1988). *Molecular dynamics simulation of despentapeptide insulin in a crystalline environment*. *J. Mol. Biol.* **200**, 571–577.

Smith, P. E. & van Gunsteren, W. F. (1995). *Reaction field effects on the simulated properties of liquid water*. *Mol. Simul.* **15**, 233–245.

Torda, A. E., Brunne, R. M., Huber, T., Kessler, H. & van Gunsteren, W. F. (1993). *Structure refinement using time-averaged J-coupling constant restraints*. *J. Biomol. NMR*, **3**, 55–66.

Torda, A. E., Scheek, R. M. & van Gunsteren, W. F. (1990). *Time-averaged nuclear Overhauser effect distance restraints applied to tendamistat*. *J. Mol. Biol.* **214**, 223–235.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). *Structure of ubiquitin refined at 1.8 Å resolution*. *J. Mol. Biol.* **194**, 531–544.

International Tables for Crystallography (2012). Vol. F, Chapter 20.2, pp. 642–648.

Chapter 20.2. Molecular-dynamics simulations of biological macromolecules

C. B. Post and V. M. Dadarlat

20.2.1. Introduction

Molecular dynamics (MD) is the simulation of motion for a system of particles. Advances in the theory of atomic interactions and the increasing availability of high-power computers have led to rapid development of this field and greater understanding of macromolecular motions. In the earliest molecular-dynamics simulations of protein molecules (McCammon *et al.*, 1977; McCammon & Harvey, 1987), the systems were greatly simplified in order to fit within the computing capabilities of that time. Simplifications included the exclusion of water molecules and even of explicit hydrogen atoms; the effect of hydrogen atoms was built into the heavy-atom properties using so-called extended-atom parameters. Simulation time periods were limited to tens of picoseconds for systems of less than 10^3 atoms. Modern simulations, by contrast, are based on improved force fields (MacKerell *et al.*, 1998) and benefit from considerable development in algorithms. In addition, the possible size and time period of simulations have increased by orders of magnitude; large systems of the order of 10^4 atoms (including explicit solvent molecules) and nanosecond time periods are accessible. With dedicated computer time, the microsecond regime is possible (Duan & Kollman, 1998). Interestingly, the first 100 ps simulation of an enzyme complex was of hen egg-white lysozyme (Post *et al.*, 1986), the first enzyme whose structure was solved by X-ray crystallography. Then the simulation required several months of dedicated time on a Cray supercomputer, but now it can be accomplished in less than a week on a common workstation.

A consequence of this enormous growth in computing power has been the particularly successful application of molecular dynamics of biological molecules to three-dimensional structure determination and refinement. It is now practical to use molecular dynamics, in combination with crystallographic and NMR data, to search the large conformational space of proteins and nucleic acids to find structures consistent with the data and to improve the agreement with the data. The advantages of molecular dynamics over manual rebuilding and least-squares refinement are the abilities to overcome the local minimum problem in an automated fashion and to search the complex conformational space of a macromolecule more extensively (Brünger *et al.*, 1987).

20.2.2. The simulation method

Molecular mechanics, whereby the energy of the system is expressed in classical terms as a function of atomic coordinates, is well established as a useful approach for describing atomic interactions (Brooks *et al.*, 1988; Goodfellow & Levy, 1998). Owing to the size of proteins and nucleic acids, the potential-energy function for large biomolecules is empirically based rather than derived from quantum-mechanical calculations. The total force on each atom, $\mathbf{F}_i$, is calculated from the gradient, or the first derivative of this potential energy, with respect to the atomic coordinates. The motion of the atom resulting from the net force is described by Newton's equation of motion,

$$\mathbf{F}_i = m_i\mathbf{a}_i, \qquad (20.2.2.1)$$

where m_i is the mass of atom i and $\mathbf{a}_i$ is the acceleration. Integration of equation (20.2.2.1) gives

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \mathbf{F}_i(t)(\Delta t)^2/(2m_i), \qquad (20.2.2.2)$$

where Δt is the time step in the integration, $\mathbf{r}_i(t+\Delta t)$ is the atomic position at time $(t+\Delta t)$, given the position $\mathbf{r}_i(t)$ at time t, and $\mathbf{v}_i(t)$ is the velocity. The forces on the particles change continuously so that a numerical solution of the equation is required. The Verlet algorithm (Verlet, 1967) or a variation, Leapfrog, is commonly used.

20.2.3. Potential-energy function

20.2.3.1. *Empirical energy*

The central element of simulations is the interaction potential between atoms as a function of atomic position, $\mathbf{r}$. The success of simulations in describing the average structure of proteins and other biological features suggests that such relatively simple potential functions adequately represent proteins and nucleic acids. The empirically based components of the energy function, $E_{\rm empir}$, include geometric terms for bond lengths, bond angles and torsion angles, and non-bonding terms for steric van der Waals interactions and electrostatic interactions. A commonly used energy function is

$$E_{\rm empir} = E_{\rm geom} + E_{\rm nonb}, \qquad (20.2.3.1)$$

$$E_{\rm geom} = \sum_{\rm bonds}(1/2)k_b(b-b_{\rm eq})^2 + \sum_{\rm angles}(1/2)k_\theta(\theta-\theta_{\rm eq})^2 + \sum_{\rm torsions}(1/2)k_\varphi[1+\cos(n\varphi-\delta)], \qquad (20.2.3.2)$$

$$E_{\rm nonb} = \sum_{\rm nbpairs}\left(q_iq_j/Dr_{ij}\right) + 4\varepsilon_{ij}\left[\left(\sigma_{ij}/r_{ij}\right)^{12} - \left(\sigma_{ij}/r_{ij}\right)^6\right], \qquad (20.2.3.3)$$

where k_b, k_θ and k_φ are force constants, $b_{\rm eq}$ and $\theta_{\rm eq}$ are equilibrium values for bond lengths, b, and angles, θ, respectively, and φ is the torsion angle of periodicity n and phase δ. The non-bonded terms depend on the interatomic distance r_{ij}, the dielectric constant D, the partial atomic charge q_i, and the van der Waals parameters ε_{ij} and σ_{ij}. The bond-stretching and angle-bending contributions are represented by harmonic potentials, while the energy associated with rotation about a bond, the torsional potential, is modelled by a cosine function [equation (20.2.3.2)]. The electrostatic component of the non-bonded interactions [first term of equation (20.2.3.3)] follows Coulomb's Law, and a Lennard–Jones 6–12 potential function [second term of equation (20.2.3.3)] is used to model steric repulsion and attractive dispersion interactions. $E_{\rm nonb}$, as a sum over pairs of atoms not involved in either a bond or bond angle, requires the use of a pairwise list between atoms. The small contribution from pairs separated by a large distance allows the use of cutoff limits for this list, but at some cost in accuracy.

Initial values for the atomic coordinates and velocities are required to begin the molecular-dynamics simulation. While initial coordinates are obtained from the model built into the electron-density map, it is necessary to generate initial velocities computationally. The most common approach is to assign random

Copyright © 2012 International Union of Crystallography

values for each atom, i, consistent with the temperature chosen for the system: $3kT/2 = (1/2)\sum m_i v_i^2$.

Integration of the equation of motion [equation (20.2.2.1)] also requires specification of the time step Δt. In the case of structure determination, this choice is limited only by the numerical stability of the calculation. Too large a value for Δt results in errors in the integration, manifested by a rapid and unacceptable increase in energy. Whereas a value for Δt of 1 to 2 fs is required for accurate trajectories and strict conservation of the energy, structure-determination protocols can employ larger values and are limited only by the need for numerical stability.

Both the temperature and Δt influence the sampling rate of conformational space. Enhanced sampling increases the rate of convergence to the structure solution. Issues related to sampling are detailed in Chapter 18.2.

20.2.3.2. Particle mesh Ewald

In an MD simulation, the accurate and rapid calculation of long-range electrostatic interactions is a central issue for the correct physical representation of the system. The Coulombic potential [first term of equation (20.2.3.3)] has been used in most cases, but the Coulombic interactions must be limited for practical reasons to a subset of pair interactions so that long-distance interactions are truncated. More recently, the Ewald method has been implemented to avoid the need for truncating the nonbonded pair list while maintaining computational efficiency. The electrostatic energy is calculated using periodic boundary conditions[1] and requires a double summation over all atoms in the central unit cell, as well as their infinite number of periodic images.

The total electrostatic energy of a periodic system with a neutral unit cell containing N point charges $q_1, q_2, \ldots, q_n$ located at positions $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N$ is

$$E_{\rm elec} = \tfrac{1}{2}\sum_n{}' \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{|\mathbf{r}_{ij} + \mathbf{n}|}.$$

The sum over $|\mathbf{n}|$ takes into account all the periodic unit cells, and $\mathbf{n}$ reflects the shape of the unit cell. For a simple periodic cubic lattice, $\mathbf{n} = (n_x L, n_y L, n_z L)$, where L is the length of the unit cell, and n_x, n_y and n_z are integers. The prime indicates that in the central unit cell the terms $i = j$ are not included. The sum is convergent for very large $|\mathbf{n}|$.

An efficient way of evaluating the triple sum above is represented by the Ewald method. In this method, each point charge is surrounded by a Gaussian charge distribution of opposite sign and same integrated magnitude as the original charge:

$$\rho_i(\mathbf{r}) = -q_i \beta^3 \exp(-\beta^2 \mathbf{r}^2)/\pi^{3/2}.$$

β is inversely proportional to the width of the charge distribution and $\mathbf{r}$ is the position relative to the centre of the distribution. The result of adding this charge distribution is a screening of the electrostatic interactions, so that the interaction between neighbouring charges is effectively short-ranged. A total screened potential is calculated by summation in the real-space lattice over the initial point-charge distribution together with the screening-charge distribution. To end up with the original point-charge distribution, a cancelling distribution having the opposite sign and same functional form as the Gaussian screening distribution must be added. Together, these Gaussian distributions form a smooth varying function of $\mathbf{r}$, which can be approximated by a superposition of continuous functions. The contribution of the cancelling distribution is calculated by adding the Fourier transforms of the distributions in reciprocal space at each point charge and transforming the total back into real space.

The final Ewald expression for the total electrostatic energy contains a real-space sum plus a reciprocal-space sum, minus a self-term introduced by the interaction of the cancelling distribution with itself:

$$E_{\rm elec} = \tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\sum_n q_i q_i \frac{\operatorname{erfc}(\beta|\mathbf{r}_{ij} + \mathbf{n}|)}{|\mathbf{r}_{ij} + \mathbf{n}|} + q_i q_j \Phi_{\rm rec}(\beta \mathbf{r}_{ij})\right] - \frac{\beta}{\pi^{1/2}}\sum_{i=1}^{N} q_i^2 + J(\mathbf{D}, \mathbf{P}, e').$$

$J(\mathbf{D}, \mathbf{P}, e')$ (de Leeuw *et al.*, 1980) depends on the total dipole moment of the unit cell, the shape of the macroscopic crystal lattice and the dielectric constant of the surrounding medium. The term $\operatorname{erfc}(\beta|\mathbf{r}_{ij} + \mathbf{n}|)$ is the complementary error function which falls to zero as $(\beta|\mathbf{r}_{ij} + \mathbf{n}|)$ increases. The term $\Phi_{\rm rec}(\beta \mathbf{r}_{ij})$ is the reciprocal Ewald sum:

$$\Phi_{\rm rec}(\beta \mathbf{r}_{ij}) = \frac{1}{\pi V}\sum_{m \neq 0} \frac{\exp(-\pi^2 \mathbf{m}^2/\beta^2)}{\mathbf{m}^2} \exp(2\pi i \mathbf{m}\mathbf{r}).$$

V is the volume and $\mathbf{m}$ is a vector in the reciprocal space. The Ewald sum in reciprocal space is the solution to Poisson's equation, with Gaussian charge densities as sources and with periodic boundary conditions. For large β, the only contribution to the sum in the real space comes from the minimum image terms. When β is large, the charge distributions are sharp and the summation in reciprocal space must be done over a large number of reciprocal-space vectors to achieve convergence.

Calculation of the Ewald sum in the original form is slow for large macromolecular solution simulations. The usual implementation of the Ewald summation for calculating electrostatic interaction is an order N^2 algorithm, or at best, $N^{3/2}$ (Christiansen *et al.*, 1993) by adjusting β to optimize computational effort. The particle mesh Ewald (PME) method (Essmann *et al.*, 1995; Darden *et al.*, 1993) is an approximation of the Ewald sum that allows an order $N \log N$ algorithm for the calculation of the total electrostatic energy. This reduction in the number of steps is accomplished by choosing β large enough that atom pairs separated by distances larger than a specified cutoff (9–12 Å) make a negligible contribution to the direct-space sum, which thus becomes an order N computation. Moreover, the reciprocal-space sum is expressed in terms of the electrostatic structure factors. Each atomic charge distribution is approximated by a gridded charge distribution. The resulting approximate structure factors are calculated by a three-dimensional fast Fourier transform applied to the grid. Using an optimized β and grid density for each simulated system, the PME method can compute the electrostatic energy for a periodic boundary system at the same level of accuracy as the classical Ewald summation in a much shorter time (Darden *et al.*, 1993).

20.2.3.3. Experimental restraints in the energy function

For the purpose of structure determination, the potential-energy function used for molecular-dynamics calculation incorporates the information from experimental data in the form of non-physical restraint terms. These restraint terms, introduced to bias the conformational sampling toward structures consistent with the experiment, are used in addition to the total potential function and, in some sense, can fulfil the requirements of a physical term in equation (20.2.2.2) (see below). The experimentally based restraint terms are added to the potential-energy function to give a total effective potential, $E_{\rm tot} = E_{\rm empir} + E_{\rm rest}$. Whereas structure-determination protocols based on NMR data employ a number of types of restraint terms, data from X-ray crystallography provide a single restraint term, $E_{\rm rest} = wE_{\rm Xray}$, the residual between the observed and calculated structure-factor amplitudes; where

---

[1] The periodic boundary conditions method allows for the simulation of the atoms in a central unit cell in the field of image unit cells generated by symmetry operations.

$$wE_{\rm Xray} = w\sum_{hkl}\left(|F_o| - k|F_c|\right)^2,$$

where w and k are scale factors.

20.2.4. Empirical parameterization of the force field

Considerable effort has gone into the development of a number of force fields for use in molecular-dynamics simulations of biomolecules (Jorgensen & Tirado-Rives, 1988; MacKerell *et al.*, 1995, 1998; van Gunsteren *et al.*, 1996). The parameters described here are those of the *CHARMM*22 force field, the force field used in the *X-PLOR* program. Estimation of the force constants, equilibrium values and non-bonding parameters in equations (20.2.3.2) and (20.2.3.3) involves a self-consistent approach that balances the bonding and non-bonding interaction terms among the macromolecule and solvent molecules (MacKerell *et al.*, 1998). A wide range of data are taken into account during an interactive process of optimization in order to adequately account for the extensive and correlated nature of the parameters in a consistent fashion. Small-molecule model compounds representative of proteins or nucleic acids are considered in detail, and a hierarchical approach is applied to extend the parameters to larger molecules with minimal adjustment at the points of connection.

The empirical basis of the parameters is broad. Gas-phase geometries and crystal structures are used to determine equilibrium bond lengths, bond angles, and dihedral phase and periodicity. Vibrational spectra, primarily from gas-phase infrared and Raman spectroscopy, are used to fit values for the force constants. Torsion-angle terms are estimated from relative energies of different conformers of model compounds, such as 4-ethylimidazole and ethylbenzene, based on gas-phase data. In cases where no satisfactory experimental data are available, *ab initio* calculations are used to obtain the required energy surfaces. Adjustments are made to describe the energy barriers and positions of saddle points, as well as the minimum-energy structures.

Optimization of the non-bonded parameters includes fitting the van der Waals and electrostatic terms of equation (20.2.3.3), while maintaining a balance among the protein–protein, water–water and protein–water interactions. The parameterization of the *CHARMM*22 force field is based on the water model and water–water interactions of the *TIP3P* model (Jorgensen *et al.*, 1983). As such, use of this parameter set with another water model will lead to inconsistencies in the balance of intermolecular interactions. Data from dipole moments, heats and free energies of vaporization, solvation and sublimation, and molecular volumes, as well as *ab initio* calculations of interaction energies and geometries are used to optimize intermolecular interactions. Partial charges of atoms are determined by fitting *ab initio* interaction energies and geometries of small-molecule compounds that model the peptide backbone and amino-acid side chains. Magnitudes and directions of dipole-moment values are also used to optimize partial charges. Experimental gas-phase dipole-moment values are used when available, while *ab initio* calculated values are adopted otherwise. The van der Waals parameters are then refined by comparing results of condensed-phase simulations on pure solvents with heats of vaporization and molecular volumes.

The crystallographic restraint term in the potential-energy function, $E_{\rm rest}$, must also be parameterized to optimize the agreement with the experimental structure-factor amplitudes while simultaneously retaining good geometry and non-bonding interactions. Optimization of $E_{\rm rest} = wE_{\rm Xray}$ involves only the estimation of w. Unlike the parameters in $E_{\rm empir}$, w has no physical basis and is usually chosen to make the force due to $E_{\rm rest}$ balance the total force contributed by all terms in $E_{\rm empir}$. As refinement of the structure progresses, these forces, and hence w, necessarily change since the quality, in terms of geometry and non-bonding interactions, of the structure improves and the crystallographic residual is reduced.

20.2.5. Modifications in the force field for structure determination

Simulated-annealing protocols require modification of the parameters to maintain the correct geometry and local structural integrity of the molecule in order to allow heating to very high, non-physical temperatures for several thousand integration steps. Such modifications are acceptable in the case of structure determination since the primary goal is to define the optimum equilibrium structure in best agreement with the crystallographic or NMR data. Simulations intended to reproduce the fluctuations or dynamic properties of the system must employ carefully defined parameters without such modifications. These modifications include substantial increases in the force constants for bond lengths and angles, *e.g.*, factors of two to ten are used in the parameters specified in the *X-PLOR* file parallhdg.pro. A number of improper torsional terms are added to maintain proper chirality.

The specific terms in $E_{\rm nonb}$ are also modified for the purpose of structure determination. In this methodology, the goal is to converge efficiently to a model that satisfies the experimental data, rather than to obtain an accurate description of the conformational surface, such as estimating fluctuations and equilibrium distributions. Alterations in $E_{\rm nonb}$ include the replacement of the computationally expensive $E_{\rm vdW}$ by a quartic or harmonic repulsive term, which prevents steric conflict among atoms, but ignores dispersive attraction. The electrostatic term, $E_{\rm elec}$, is frequently excluded altogether, since the $1/\mathbf{r}$ dependence of the Coulombic potential allows charge interactions to dominate the interatomic forces far from the global minimum in a fashion that hinders movement toward the global minimum. Exclusion of this important physical property of biological systems is possible, because the crystallographic structure factors contain sufficient information to reflect adequately the imprint of electrostatics on the average structure.

20.2.6. Internal dynamics and average structures

It is most often the goal of the structural biologist to define a single average structure of a macromolecule. The well recognized internal motions arising from thermal fluctuations of a macromolecule may be necessary for function, but, nonetheless, the methods of structure determination generally aim to model a single average structure. Internal motions range from the high frequency, small amplitude motions (*i.e.* those modelled by crystallographic B values) to low frequency, larger amplitude motions of loops and whole domains. Some studies (Kuriyan *et al.*, 1986; Post, 1992) have examined the validity of the assumptions about fast timescale motions made by the methods of structure determination. It is reasonable that some of the differences between the structure solutions of a protein obtained by NMR spectroscopy and X-ray crystallography are due to differences in the effects of internal motions. The application of molecular-dynamics algorithms for structure determination has allowed the use of protocols that account for effects of internal motions by employing time-averaged restraints (Schiffer *et al.*, 1995).

20.2.7. Assessment of the simulation procedure

Equation (20.2.3.1) is a reasonable representation of the energy function of proteins. This point is illustrated here with results from 0.8 ns molecular-dynamics simulations of hen egg-white lysozyme (PDB entry 1lzt, 1.97 Å resolution), bovine pancreas ribonuclease A (5rsa, 2.0 Å resolution), bovine α-lactalbumin (1hfz, 2.3 Å resolution) and trypsin (2ptn, 1.55 Å resolution). [The coordinates for bovine α-lactalbumin were kindly provided by K. R. Acharya prior to their publication (Pike *et al.*, 1996).]

The proteins were fully hydrated, and the simulations were calculated with the *CHARMM*22 force field, using truncated octahedron periodic boundary conditions. The four proteins were overlaid with bulk water from an equilibrated simulation, and a 10 ps trajectory was calculated for the rearrangement of the water molecules around the protein with fixed protein atoms. The number of water molecules (4000 to 6000) required to hydrate the protein varied with the protein size (123–223 residues) and shape, with at least four layers of water molecules between the peripheral protein atoms and the walls of the boxes. The simulations were performed at constant pressure and temperature ($T = 300$ K and $p = 1$ atm) using the extended-system algorithms (Hoover, 1985; Nose, 1984) implemented in *CHARMM*. The 300 K constant temperature was maintained by coupling to an external bath, with a coupling constant of 25 ps. The *SHAKE* algorithm (Ryckaert *et al.*, 1977) was used to constrain bond lengths between hydrogen atoms and heavy atoms, allowing for a time step of 2 fs in the integration of the equations of motion. A non-bonded cutoff of 12 Å was used for the Lennard–Jones potential calculation. The electrostatic forces and energies were computed using the PME method (Darden *et al.*, 1993; Essmann *et al.*, 1995). The PME charge grid spacing was 0.7 Å, and the charge grid was interpolated with the direct sum tolerance set to 4.0×10^{-6}. The non-bonded pair lists were updated every 10 steps. Structures for analysis were saved every 0.1 ps.

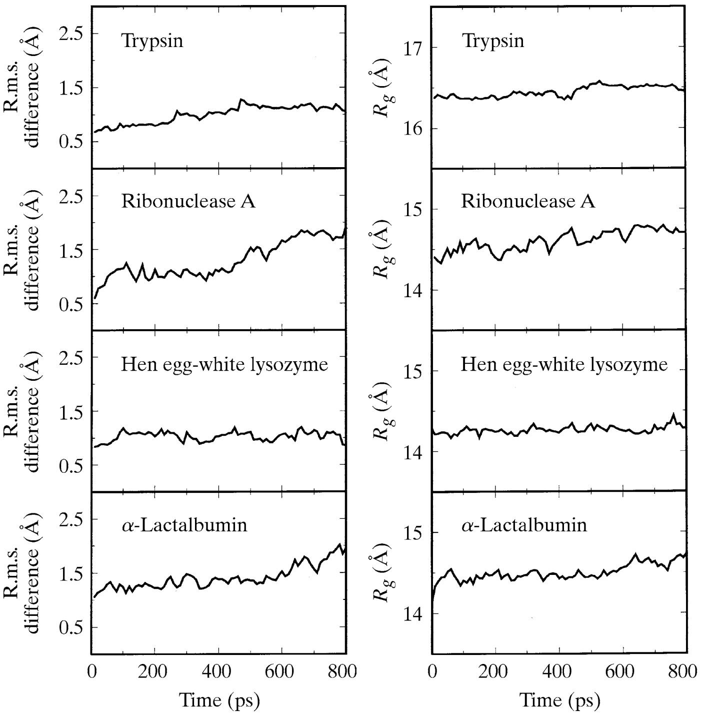


Figure 20.2.7.1
Structural comparison and radii of gyration of various proteins as a function of time in the molecular-dynamics simulation. Left: r.m.s. coordinate differences averaged over main-chain atoms (N, Cα, C) between the energy-minimized crystallographic structure and the simulation snapshot. Right: radii of gyration (R_g).

Fig. 20.2.7.1 shows the deviation during the simulation period of the main-chain coordinates between the simulation structures and the crystallographic starting structure. The r.m.s. coordinate deviations in the case of lysozyme are particularly stable and small: approximately 1.0 Å over the course of the trajectory. Other r.m.s. values are more typical and range from 1.0 to 2.0 Å. The time dependence of other properties, such as the radius of gyration, can also be used to follow the stability and behaviour of a trajectory. These time series, also shown in Fig. 20.2.7.1, are constant. Jumps in such a time series can be used to detect conformational transitions (Post *et al.*, 1989).

Other experimental properties have been compared in the literature with those calculated from molecular-dynamics trajectories. Of particular interest is comparing time-dependent properties measured by NMR spectroscopy. An approach to calculating NMR relaxation rates was recognized early on when development of both the molecular-dynamics simulations of proteins and a model-independent theory for NMR relaxation was started (Levy, Karplus & McCammon, 1981; Levy, Karplus & Wolynes, 1981; Lipari & Szabo, 1982). Since then, the common practice of isotopic labelling of proteins for NMR structure determination has allowed the measurement of numerous NMR relaxation rates, particularly rates that characterize the motion of backbone atoms. Long simulations have been conducted to compare the calculated and experimental values (Abseher *et al.*, 1995; Chatfield *et al.*, 1998; Smith *et al.*, 1995). In a particularly long simulation, an 11 ns trajectory period was used to estimate relaxation rates associated with the motions of the vectors N—H, Cα—H and C—H methyl groups from alanine and leucine (Chatfield *et al.*, 1998). Trends in the general order of mobility of these vectors are reproduced, although a residue-by-residue comparison shows some differences.

20.2.8. Effect of crystallographic atomic resolution on structural stability during molecular dynamics

The variation in r.m.s. deviation between the initial crystallographic structure and the simulation coordinates for different protein trajectories (Fig. 20.2.7.1) raises the question of whether the atomic resolution of the starting X-ray structure influences the magnitude of this deviation. In order to investigate this issue, we calculated trajectories for bovine pancreatic trypsin inhibitor (BPTI), starting with crystallographic structures determined from data at three different atomic resolutions: 1bpi at 1.1 Å resolution (Parkin *et al.*, 1999), 6pti at 1.7 Å resolution (Wlodawer *et al.*, 1987) and 1bhc at 2.7 Å resolution (Hamiaux *et al.*, 1999). The errors in the atomic coordinates estimated from the Luzzati plots are 0.06 Å for the 1.1 Å resolution structure and 0.41 Å for the 2.7 Å resolution structure. The protocol described in the previous

Table 20.2.8.1
R.m.s. coordinate differences between crystallographic structures and average MD structures

The upper half of the matrix contains values for main-chain atoms, while the lower half contains values for side-chain heavy atoms.

		X-ray structure			Energy-minimized X-ray structure			Average MD structure		
		1.1 Å	1.7 Å	2.7 Å	1.1 Å	1.7 Å	2.7 Å	1.1 Å	1.7 Å	2.7 Å
X-ray structure	1.1 Å		0.54	0.51	0.4	0.61	0.7	0.97	1.09	1.13
	1.7 Å	1.5		0.41	0.6	0.42	0.57	1.11	1.14	1.11
	2.7 Å	1.62	1.43		0.59	0.57	0.48	1.12	1.12	1.11
Energy-minimized X-ray structure	1.1 Å	0.57	2.81	2.06		0.59	0.63	1.01	1.15	1.13
	1.7 Å	1.5	0.79	1.68	1.47		0.58	1.12	1.13	1.13
	2.7 Å	1.72	1.65	0.99	1.78	1.77		1.28	1.11	1.11
Average MD structure	1.1 Å	1.89	2.18	2.45	1.94	2.87	2.66		1.06	1.08
	1.7 Å	2.53	2.28	2.49	2.52	2.33	2.52	2.22		0.87
	2.7 Å	2.15	2.05	1.9	2.17	2.13	1.85	2.17	2.15	

section was followed for simulations starting with each of the three crystallographic structures over a 500 ps simulation time. The net charge of +6 e on BPTI was neutralized by adding six chloride anions to the solvated protein system, thus accomplishing the ideal conditions for a PME calculation for the electrostatic interaction. The truncated octahedra contain approximately 3700 water molecules, and the total number of atoms in the simulations is over 12 000. The simulations were carried out on an eight-node IBM/SP2 and required 4.5 h of CPU time per 10 ps of dynamics run.

Root-mean-square differences (r.m.s.d.'s) in atomic coordinates were calculated between all pairs of coordinates from the X-ray structures, the energy-minimized X-ray structures and the 10 ps average MD structure obtained near 300 ps of the simulation period. In Table 20.2.8.1, the upper diagonal r.m.s.d. values are the main-chain-atom differences, while the lower diagonal ones are the side-chain-atom differences. The r.m.s.d.'s between the three X-ray structures range from 0.4–0.5 Å for the main-chain atoms and 1.4–1.6 Å for the side-chain atoms. The larger r.m.s. values for averages over side-chain atoms imply alternative side-chain orientations. This degree of structural deviation is visualized with an overlay of the three X-ray structures of BPTI shown on the left-hand side of Fig. 20.2.8.1. Energy minimization with respect to the *CHARMM* force field alters the main-chain atoms of the X-ray structures by approximately 0.4 Å and increases the differences between two X-ray structures to 0.6–0.7 Å. The differences in the main-chain atoms between an MD average structure and an energy-minimized X-ray structure are only somewhat larger: 1.0–1.1 Å. Interestingly, these values are comparable to the values obtained when comparing the three MD average structures: 0.9–1.1 Å. The main-chain structural differences among the three 10 ps average MD structures are shown on the right-hand side of Fig. 20.2.8.1. The general trends observed for main-chain atoms are also found for side-chain atoms. Thus, the differences between the X-ray structures increase somewhat as a result of energy minimization, and the differences between MD average structures and X-ray structures (1.9–2.9 Å) are similar to those between two X-ray structures (1.5–2.8 Å) or two MD average structures (approximately 2.2 Å).

Similar r.m.s. values are found if the starting velocites for a simulation are varied while maintaining the same starting coordinates (Caves *et al.*, 1998); the r.m.s.d.'s obtained from 120 ps MD simulations were 0.7–1.1 Å for the main-chain atoms with respect to the reference X-ray structure and 0.8–1.5 Å between MD individual trajectory averages. The results given in Table

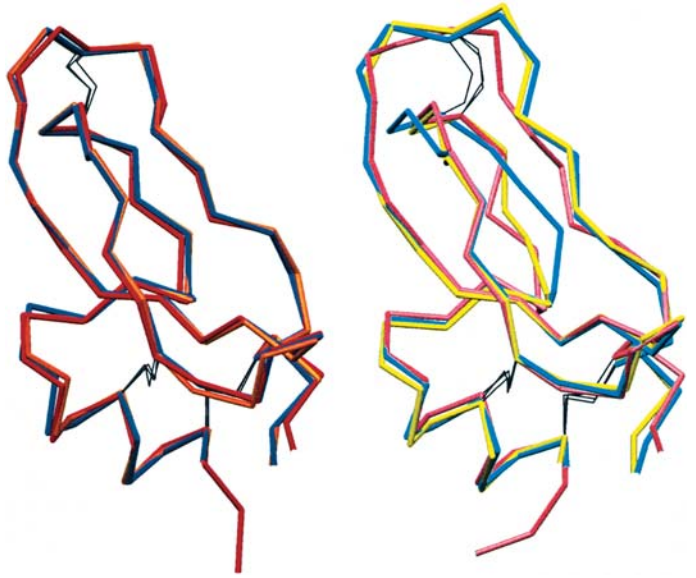


Figure 20.2.8.1
Cα tracings of BPTI. Left: crystallographic structures determined from data at three different resolutions: 1.1 (red), 1.7 (blue) and 2.7 Å (orange). Right: 10 ps average MD structures from simulations initiated with the energy-minimized crystallographic structure determined at 1.1 (pink), 1.7 (cyan) or 2.7 Å (yellow) resolution. The 10 ps average is over coordinates from 290 to 300 ps.

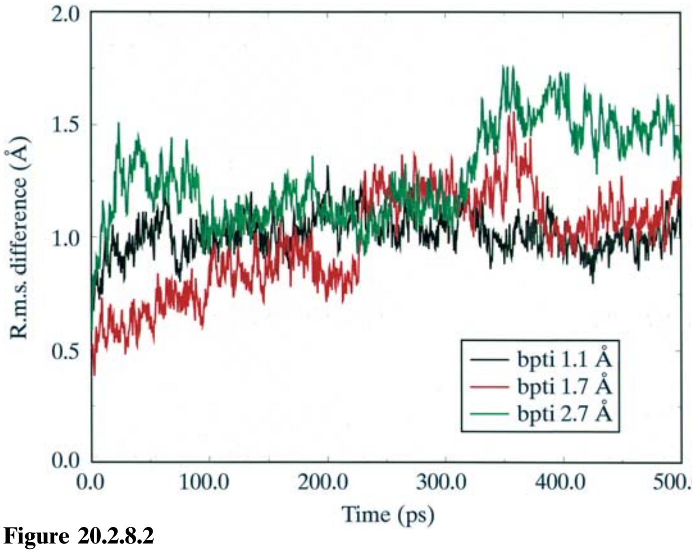


Figure 20.2.8.2
BPTI r.m.s. coordinate differences between the energy-minimized crystallographic structure and MD snapshots from three simulations. A simulation was initiated from the energy-minimized crystallographic structure determined at 1.1 (black), 1.7 (red) or 2.7 Å (green) resolution. The r.m.s.d. is averaged over the main-chain atoms N, Cα and C.

20.2.8.1, together with those of Caves *et al.* (1998), suggest that sampling on the nanosecond timescale largely reflects the conformational variations due to thermal fluctuations that result from a potential-energy surface with multiple minima separated by low barriers (Cooper, 1976). In this context, MD simulations starting with different X-ray structures offer a more extensive sampling of the conformational space of the specific protein than simulations carried out from a single X-ray structure, although this conclusion remains to be demonstrated by a more thorough analysis. Our results do not support the conclusion that overall r.m.s.d.'s between MD average structures and the starting X-ray structures correlate with atomic resolution.

The r.m.s.d.'s between main-chain atoms in the starting X-ray structures and simulation snapshots as a function of time are presented in Fig. 20.2.8.2. The 1.1 Å resolution structure has the most stable trajectory during the 500 ps trajectory, with an average r.m.s. value of 1.01 (9) Å. The 1.7 Å resolution structure has an r.m.s. value of 0.98 (22) Å. In this simulation, the r.m.s.d.'s fluctuate more widely from the average value, with small differences in the first 200 ps, larger ones between 200 and 400 ps, and again smaller ones in the last 100 ps. For the 2.7 Å resolution structure, the average over the simulation is 1.28 (21) Å. From the results presented here, it is concluded that the higher-resolution structures are more stable during MD simulations and have a shorter equilibration period (50 ps for 1.1 Å resolution and over 300 ps for 2.7 Å resolution). This conclusion is consistent with larger errors in the atomic coordinates of X-ray structures determined from lower-resolution data.

This work was supported by grants to CBP from the NIH (R01-GM39478, AI39639). CBP was supported by a Research Career Development Award from the NIH (K04-GM00661) and VMD is a DOE/SLOAN Postdoctoral Fellow in computational biology. The computing facilities shared by the Structural Biology group were supported by grants from the Lucille P. Markey Foundation and the Purdue University Academic Reinvestment Program.

References

Abseher, R., Ludemann, S., Schreiber, H. & Steinhauser, O. (1995). *NMR cross-relaxation investigated by molecular dynamics simulation: a case study of ubiquitin in solution. J. Mol. Biol.* **249**, 604–624.

Brooks, C. L. III, Karplus, M. & Pettitt, B. M. (1988). *Proteins: a theoretical perspective of dynamics, structure and thermodynamics. Adv. Chem. Phys.* **71**, 1–259.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Crystallographic R factor refinement by molecular dynamics. Science*, **235**, 458–460.

Caves, L. S. D., Evanseck, J. D. & Karplus, M. (1998). *Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. Protein Sci.* **7**, 649–666.

Chatfield, D. C., Szabo, A. & Brooks, B. R. (1998). *Molecular dynamics of staphylcoccal nuclease: comparison of simulation with 15N and 13C NMR relaxation data. J. Am. Chem. Soc.* **120**, 5301–5311.

Christiansen, D., Peram, J. W. & Petersen, H. G. (1993). *On the fast multipole method for computing the energy of periodic assemblies of charged and dipolar particles. J. Comput. Phys.* **107**, 403–405.

Cooper, A. (1976). *Thermodynamic fluctuations in protein molecules. Proc. Natl Acad. Sci. USA*, **92**, 2740–2741.

Darden, T., York, D. & Pedersen, L. (1993). *Particle mesh Ewald (PME): a N log(N) method for Ewald sums in large systems. J. Chem. Phys.* **98**, 10089–10092.

Duan, Y. & Kollman, P. A. (1998). *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science*, **282**, 740–744.

Essmann, U., Perrera, L., Berkovitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995). *A smooth particle mesh Ewald method. J. Chem. Phys.* **103**, 8577–8593.

Goodfellow, J. M. & Levy, R. M. (1998). *Theory and simulation. Curr. Opin. Struct. Biol.* **8**, 209–210.

Gunsteren, W. F. van, Billeter, S. R., Eising, A. A., Hünenberger, P. H., Mark, A. E., Scott, W. R. P. & Tironi, I. G. (1996). *Biomolecular Simulation: the GROMOS96 Manual and User Guide*. Vdf Hochschulverlag, Zurich, Switzerland.

Hamiaux, C., Prangé, T., Riès-Kautt, M., Ducruix, A., Lafont, S., Astier, J. P. & Veesler, S. (1999). *The decameric structure of bovine pancreatic trypsin inhibitor (BPTI) crystallized from thiocyanate at 2.7 Å resolution. Acta Cryst.* D**55**, 103–113.

Hoover, W. G. (1985). *Canonical dynamics: equilibrium phase-space distributions. Phys. Rev. A*, **31**, 1695–1697.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *Comparison of simple potential functions for simulating liquid water. J. Chem. Phys.* **79**, 926–935.

Jorgensen, W. L. & Tirado-Rives, J. (1988). *The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc.* **110**, 1657–1666.

Kuriyan, J., Petsko, G. A., Levy, R. M. & Karplus, M. (1986). *Effect of anisotropy and anharmonicity on protein crystallographic refinement: an evaluation by molecular dynamics. J. Mol. Biol.* **190**, 455–479.

Leeuw, S. W. de, Perram, J. W. & Smith, E. R. (1980). *Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants. Proc. R. Soc. London Ser. A*, **373**, 27–56.

Levy, R. M., Karplus, M. & McCammon, J. A. (1981). *Increase of ^{13}C NMR relaxation times in proteins due to picosecond motional averaging. J. Am. Chem. Soc.* **103**, 994–996.

Levy, R. M., Karplus, M. & Wolynes, P. G. (1981). *NMR relaxtion parameters in molecules with internal motion: exact Langevin trajectory results compared with simplified relaxation models. J. Am. Chem. Soc.* **103**, 5998–6011.

Lipari, G. & Szabo, A. (1982). *Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. J. Am. Chem. Soc.* **104**, 4546–4559.

McCammon, J. A., Gelin, B. R. & Karplus, M. (1977). *Dynamics of folded proteins. Nature (London)*, **267**, 585–590.

McCammon, J. A. & Harvey, S. C. (1987). *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press.

MacKerell, A., Wiorkiewicz-Kuczera, J. & Karplus, M. (1995). *An all atom empirical energy function for the simulation of nucleic acids. J. Am. Chem. Soc.* **117**, 11946–11975.

MacKerell, A. D. Jr, Bashford, D., Bellott, M., Dunbrack, R. L. Jr, Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E. III, Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J. & Karplus, M. (1998). *All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B*, **102**, 3586–3616.

Nose, S. (1984). *A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys.* **81**, 511–519.

Parkin, S., Rupp, B. & Hope, H. (1999). *The structure of bovine pancreatic trypsin inhibitor at 125 K: definition of carboxyl-terminal residues glycine-57 and alanine 58*. In preparation.

Pike, A. C., Brew, K. & Acharya, K. R. (1996). *Crystal structures of guinea-pig, goat and bovine α-lactalbumin highlight the enhanced conformational flexibility of regions that are significant for its action in lactose synthase. Structure*, **4**, 691–703.

Post, C. B. (1992). *Internal motional averaging and three-dimensional structure determination by NMR. J. Mol. Biol.* **224**, 1087–1101.

Post, C. B., Brooks, B. R., Karplus, M., Dobson, C. M., Artymiuk, P. J., Cheetham, J. C. & Phillips, D. C. (1986). *Molecular dynamics simulations of native and substrate-bound lysozyme. J. Mol. Biol.* **190**, 455–479.

Post, C. B., Dobson, C. M. & Karplus, M. (1989). *A molecular dynamics analysis of protein structural elements. Proteins*, **5**, 337–354.

Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. (1977). *Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys.* **23**, 327–341.

Schiffer, C. A., Gros, P. & van Gunsteren, W. F. (1995). *Time-averaging crystallographic refinement: possibilities and limitations using α-cyclodextrin as a test system. Acta Cryst.* D**51**, 85–92.

Smith, P. E., van Schaik, R. C., Szyperski, T., Wuthrich, K. & van Gunsteren, W. F. (1995). *Internal mobility of the basic pancreatic trypsin inhibitor in solution: a comparison of NMR spin relaxation measurements and molecular dynamics simulation. J. Mol. Biol.* **246**, 356–365.

Verlet, L. (1967). *Computer experiments on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules. Phys. Rev.* **159**, 98–103.

Wlodawer, A., Nachman, J., Gilliland, G. L., Gallagher, W. & Woodward, C. (1987). *Structure of form III crystals of bovine pancreatic trypsin inhibitor. J. Mol. Biol.* **198**, 469–480.

International Tables for Crystallography (2012). Vol. F, Chapter 21.1, pp. 649–661.

PART 21. STRUCTURE VALIDATION

Chapter 21.1. Validation of protein crystal structures

G. J. Kleywegt

21.1.1. Introduction

Owing to the limited resolution and imperfect phase information that macromolecular crystallographers usually have to deal with, building and refining a protein model based on crystallographic data is not an exact science. Rather, it is a subjective process, governed by experience, prejudices, expectations and local practices (Brändén & Jones, 1990; Kleywegt & Jones, 1995*b*, 1997). This means that errors in this process are almost unavoidable, but it is the crystallographer's task to remove as many of these as possible prior to analysis, publication and deposition of the structure. With high-resolution data and good phases, the resulting model is probably more than 95% a consequence of the data, although even at atomic resolution, subjective choices must still be made: which refinement program to use, whether to include alternative conformations, whether to model explicit H atoms, how to model temperature factors, which restraints and constraints to apply, which peaks in the maps to interpret as solvent molecules and how to treat noncrystallographic symmetry (NCS). Once the resolution becomes worse than ~2 Å, this balance shifts and some published protein models appear to have been determined more by some crystallographer's imagination than by any experimental data.

Subjectivity is not necessarily a problem, provided that the crystallographer is experienced, knows what he or she is doing and is aware of the limitations that the experimental data impose on the model. However, even inexperienced people can avoid many of the pitfalls of model building and refinement. Supervisors have a major responsibility in this respect: education is an important factor (Dodson *et al.*, 1996). Students who have built and refined a previously determined structure from scratch as a training exercise will have met most of the problems that can be encountered in real life (Jones & Kjeldgaard, 1997). Apart from hands-on experience, there are many other methods to reduce or avoid errors. These include (1) the use of information derived from databases of well refined structures in model building (Kleywegt & Jones, 1998) [*e.g.* to generate main-chain coordinates from a C$^{\alpha}$ trace (Jones & Thirup, 1986) and side-chain coordinates from preferred rotamer conformations (Ponder & Richards, 1987)]; (2) the use of various sorts of local quality checks (to detect residues that for one or more reasons are deemed 'unusual' and that require further scrutiny and perhaps adjustment; Kleywegt & Jones, 1996*a*, 1997); and (3) the use of global quality indicators [*e.g.* the use of the free *R* value (Brünger, 1992*a*, 1993) to signal major errors, to prevent overfitting, and to monitor the progress of the rebuilding and refinement process (Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996; Brünger, 1997)].

21.1.2. Types of error

At every step of a crystal structure determination, the danger of making mistakes looms (Brändén & Jones, 1990; Janin, 1990). In this laboratory, for instance, a protein other than that intended was once purified (lysozyme instead of cellular retinoic acid binding protein), which obviously made the molecular-replacement problem rather intractable. Similarly, there is at least one published crystallization report of a protein other than for which the crystallographers had hoped: crystals of the light-harvesting complex LH1 actually turned out to be of bacterioferritin (Nunn *et al.*, 1995). There is at least one case in which an incorrect molecular-replacement solution was found that persisted all the way to the final published model, namely that of turkey egg-white lysozyme (Bott & Sarma, 1976). During data collection and processing, space-group assignment errors are occasionally made, such as in the case of chloromuconate cycloisomerase (Hoier *et al.*, 1994; Kleywegt *et al.*, 1996). A more common problem at this stage, however, is weak and/or incomplete data. The importance of complete data sets with a high signal-to-noise ratio and high redundancy for the success of the subsequent structure determination process (phasing, model building and refinement) cannot be overstressed. However, the discussion in this chapter will focus mainly on errors that may creep into a protein model built and refined by a crystallographer. Such errors come in various classes (Brändén & Jones, 1990) and, fortunately, the frequency of each type of error is inversely proportional to its seriousness.

1. In the worst case, the model (or a sub-unit) may essentially be completely wrong. Recently identified examples of this type of problem include asparaginase/glutaminase (Ammon *et al.*, 1988; Lubkowski *et al.*, 1994) and photoactive yellow protein (McRee *et al.*, 1989; Borgstahl *et al.*, 1995).
2. In other cases, secondary-structure elements may have been correctly identified for the most part, but incorrectly connected. This happened, for instance, in the structure determination of D-Ala-D-Ala carboxypeptidase/transpeptidase (Kelly *et al.*, 1986; Kelly & Kuzin, 1995).
3. A fairly common mistake during the initial tracing is to overlook one residue, which leads to a register error (or frame shift). The model is usually brought back into register with the density a bit further down the sequence, where the opposite error is made (*e.g.* an extra residue is inserted into density for a turn). This is a serious error, but it is usually possible to detect and correct it in the course of the refinement and rebuilding process (Kleywegt *et al.*, 1997). However, it is not impossible for such an error to persist, particularly in low-resolution studies. Indeed, in one case in which a published 3.0 Å structure was re-refined, a register error was detected involving about two dozen residues (Hoier *et al.*, 1994; Kleywegt *et al.*, 1996).
4. Sometimes the primary sequence used by the crystallographer contains one or more mistakes. These may arise from post-translational modifications, from sequencing errors, from the absence of a published amino-acid sequence at the time of tracing, from unanticipated cloning artifacts or simply from trivial 'transcription artifacts'. In this laboratory,

Copyright © 2012 International Union of Crystallography

the latter occurred during the refinement of human α-class glutathione S-transferase A1-1 (Sinning *et al.*, 1993), where one glycine residue had mistakenly been typed in as aspartate. Fortunately, the error revealed itself even at low resolution (2.6 Å), because the model was refined conservatively. In this case, the group of side-chain atoms obtained a very high *B* factor, in contrast to the very low *B* factor for the grouped main-chain atoms.

(5) The most common type of model-building error is locally incorrect main-chain and/or side-chain conformations. Such errors are easy to make in low-resolution maps calculated with imperfect phases. Moreover, multiple conformations are often unresolved even at moderately high resolution (~2 Å), which further complicates the interpretation of side-chain density. Nevertheless, many of them can be avoided through the use of information derived from databases (such as rotamer conformations; Jones *et al.*, 1991; Zou & Mowbray, 1994; Kleywegt & Jones, 1998) and careful rebuilding and refinement protocols (Kleywegt & Jones, 1997).

(6) Various types of error (possibly, to some extent, compensating ones) can be introduced during refinement, particularly if a satisfactorily low value for the conventional crystallographic *R* value is desired (Kleywegt & Jones, 1995*b*). This can always be achieved (even for models that have been deliberately traced backwards through the density; Jones *et al.*, 1991; Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996) by removing data that do not agree well with the model (through a resolution and σ cutoff), by not exploiting the redundancy of noncrystallographic symmetry properly (Kleywegt & Jones, 1995*b*; Kleywegt, 1996), by using an inappropriate temperature-factor model, by introducing alternative conformations and refining occupancies when these are not warranted by the information content of the data, by sprinkling the model with solvent molecules, and by reducing the weight given to the geometric and other restraints relative to the weight given to the crystallographic data.

One should realise that making errors is almost unavoidable (given the fact that one usually deals with limited resolution and less than perfect phases). The purpose of refinement and rebuilding is to detect and fix the errors to obtain the best possible final model that will be interpreted in terms of the biological role of the protein. Nevertheless, sometimes errors do persist into the publication and the deposited model. This may be a consequence of factors such as (Jones & Kjeldgaard, 1997):

(1) inexperienced, under-supervised people who do the work (and have a supervisor who may be in a hurry to publish);
(2) computer programs used as black boxes;
(3) new methods not adopted until the limitations of older ones have been experienced;
(4) intermediate models not subjected to critical and systematic quality analysis;
(5) use of 'quality indicators' that are strongly correlated with parameters that are restrained during refinement (r.m.s. deviation of bond lengths and angles from ideal values, r.m.s. ΔB for bonded atoms *etc.*).

21.1.3. Detecting outliers

21.1.3.1. Classes of quality indicators

Many statistics, methods and programs were developed in the 1990s to help identify errors in protein models. These methods generally fall into two classes: one in which only coordinates and *B* factors are considered (such methods often entail comparison of a model to information derived from structural databases) and another in which both the model and the crystallographic data are taken into account. Alternatively, one can distinguish between methods that essentially measure how well the refinement program has succeeded in imposing restraints (*e.g.* deviations from ideal geometry, conventional *R* value) and those that assess aspects of the model that are 'orthogonal' to the information used in refinement (*e.g.* free *R* value, patterns of non-bonded interactions, conformational torsion-angle distributions). An additional distinction can be made between methods that provide overall (global) statistics for a model (such methods are suitable for monitoring the progress of the refinement and rebuilding process) and those that provide information at the level of residues or atoms (such methods are more useful for detecting local problems in a model). It is important to realise that almost all coordinate-based validation methods detect *outliers* (*i.e.* atoms or residues with unusual properties): to assess whether an outlier arises from an *error* in the model or whether it is a genuine, but unusual, *feature* of the structure, one must inspect the (preferably unbiased) electron-density maps (Jones *et al.*, 1996)!

In this section, some quality indicators will be discussed that have been found to be particularly useful in daily protein crystallographic practice for the purpose of detecting problems in intermediate models. Section 21.1.7 provides a more extensive discussion of many of the quality criteria that are or have been used by macromolecular crystallographers.

21.1.3.2. Local statistics

From a practical point of view, these are the most useful for the crystallographer who is about to rebuild a model. Examples of useful quality indicators are:

(1) The real-space fit (Jones *et al.*, 1991; Chapman, 1995; Jones & Kjeldgaard, 1997; Vaguine *et al.*, 1999), expressed as an *R* value or as a correlation coefficient between 'observed' and calculated density. This property can be calculated for any subset of atoms, *e.g.* for an entire residue, for main-chain atoms or for side-chain atoms. It is best to use a map that is biased by the model as little as possible [*e.g.*, a σ_A-weighted map (Read, 1986), an NCS-averaged map (Kleywegt & Read, 1997) or an omit map (Bhat & Cohen, 1984; Hodel *et al.*, 1992)]. In practice, the real-space fit is strongly correlated with the atomic temperature factors, even though these are not used in the calculations.

(2) The Ramachandran plot (Ramakrishnan & Ramachandran, 1965; Kleywegt & Jones, 1996*b*). Residues with unusual main-chain φ, ψ torsion-angle combinations that do not have unequivocally clear electron density are almost always in error. However, one should keep in mind that the error may have its origin in (one of) the neighbouring residues. For instance, if the peptide O atom of a residue is pointing in the wrong direction, the φ value for the next residue may be off by 150–180° (Kleywegt, 1996; Kleywegt & Jones, 1998).

(3) The pep-flip value (Jones *et al.*, 1991; Kleywegt & Jones, 1998). This statistic measures the r.m.s. distance between the peptide O atom of a residue and its counterparts found in a database of well refined high-resolution structures that occur in parts of those structures with a similar local C^α backbone conformation. If the pep-flip value is large (*e.g.* >2.5 Å), the residue is termed an outlier, but whether it

is an error can only be determined by inspecting the local density.

(4) The rotamer side-chain fit value (Jones *et al.*, 1991; Kleywegt & Jones, 1998). This statistic measures the r.m.s. distance between the side-chain atoms of a residue and those in the most similar rotamer conformation for that residue type. A value greater than ~1.0–1.5 Å signals an outlier. In many cases (particularly, but not exclusively, at low resolution), a non-rotamer side chain can easily be replaced by a rotamer conformation, perhaps in conjunction with a slight rigid-body movement of the entire residue or with some adjustment of the side-chain torsion angles (Zou & Mowbray, 1994; Kleywegt & Jones, 1997).

(5) Hydrogen-bonding analysis. The correct orientation of histidine, asparagine and glutamine side chains cannot usually be inferred from electron density alone. Inexperienced crystallographers can benefit from suggestions based on the analysis of hydrogen-bonding networks (Hooft *et al.*, 1996*b*), although every case should be examined critically (*e.g.* the program does not know about solvent molecules that have not yet been added to the model or that cannot be placed because of the limitations of the data; in addition, sometimes an amino group may be interacting with an aromatic side chain).

In addition to these criteria, residues with other unusual features should be examined in the electron-density maps for the crystallographer to be able to decide whether they are in error. Such features may pertain to unusual temperature factors, unusual occupancies, unusual bond lengths or angles, unusual torsion angles or deviations from planarity (*e.g.* for the peptide plane), unusual chirality (*e.g.* for the C^α atom of every residue type except glycine), unusual differences in the temperature factors of chemically bonded atoms, unusual packing environments (Vriend & Sander, 1993), very short distances between non-bonded atoms (including symmetry mates), large positional shifts during refinement, unusual deviations from noncrystallographic symmetry (Kleywegt & Jones, 1995*b*; Kleywegt, 1996) *etc.*

21.1.3.3. Global statistics

The crystallographic *R* value used to be the major global quality indicator until it was realised that it can easily be fooled, especially at low resolution (Brändén & Jones, 1990; Jones *et al.*, 1991; Brünger, 1992*a*; Kleywegt & Jones, 1995*b*). The free *R* value, introduced by Brünger (1992*a*, 1993), has been shown to be much more reliable and harder to manipulate (Kleywegt & Brünger, 1996; Brünger, 1997). It is excellently suited for monitoring the progress of refinement, for detecting major problems with model or data and for helping reduce over-fitting of the data (which occurs if many more parameters are refined in a model than is warranted by the information content of the crystallographic data). Moreover, the free *R* value can be used to estimate the coordinate error of the final model (Kleywegt *et al.*, 1994; Kleywegt & Brünger, 1996; Brünger, 1997; Cruickshank, 1999).

In addition, the average or r.m.s. values for many of the local statistics, their minimum or maximum values or the percentage of outliers can be quoted and used to obtain an impression of the overall quality of the model and the overall fit of the model to the data.

21.1.4. Fixing errors

The object of model rebuilding is generally twofold: (1) to make the model as complete and detailed as the data will allow one to do confidently (*e.g.* to add previously unmodelled loops, ligands, water molecules *etc.*) and (2) to remove errors. At first glance, it may not seem all that important to fix each and every side chain and to correct all peptide O atoms that are pointing in the wrong direction, but one should keep in mind that an error in the scattering factor (atom type or charge), position or *B* factor of even a single atom will be detrimental to the entire model. Particularly in the early stages of model rebuilding and refinement, one often finds that after an extensive round of rebuilding followed by more refinement, the density improves dramatically and new features become clear. One should also keep in mind that incorrect features of a model may be very persistent and become 'self-fulfilling prophecies', a phenomenon known as 'model bias' (Ramachandran & Srinivasan, 1961; Read, 1986, 1994, 1997; Hodel *et al.*, 1992). This is particularly relevant in cases where unbiased phase information (*e.g.* SIRAS, MIR or MAD phases, or phases obtained after NCS or multiple-crystal averaging) is not available.

For error detection to be effective, it is best not to approach the rebuilding process in a haphazard way (Kleywegt & Jones, 1997). *O* users can employ a program called *OOPS* (Kleywegt & Jones, 1996*a*) to carry out this task in a systematic yet convenient fashion. This program uses information calculated by *O* (*e.g.* pep-flip and real-space fit values) and retrieves or derives other information from a PDB file of the current model (*e.g.* temperature factors, Ramachandran plot, changes with respect to the previous model). Moreover, results from a coordinate-based quality check by the *WHAT IF* program (Vriend, 1990; Hooft *et al.*, 1996) can be included. In all, several dozen quality indicators can be used and plots and statistics for many of these can be produced by the program. The program's most useful feature, however, is that it will generate *O* macros that when executed in *O* will take the crystallographer on a journey to all the residues that may require attention because they are outliers for one or more quality criteria. This makes the rebuilding process often faster and certainly more efficient and focused than a residue-by-residue walk through the model. In addition, it teaches inexperienced crystallographers to recognize and diagnose common model errors.

If a residue is an outlier for a certain criterion, the crystallographer has to inspect the local density and the structural context and decide the course of action. If the residue is in a region of the model in which many residues are outliers for many criteria, there may be something seriously wrong locally (for instance, there could be a register error), possibly because the density is poor. If there is poor density for several residues in a row, the crystallographer might consider leaving these residues out of the model for the next refinement round or cutting off the side chains at the C^β atoms. Sometimes local errors are correlated, such as a pep-flip error in one residue and a Ramachandran violation for its C-terminal neighbour, or a residue with a non-rotamer conformation and high temperature factors in conjunction with a poor real-space fit. *O* contains many tools to manipulate individual residues and atoms (Jones *et al.*, 1991; Jones & Kjeldgaard, 1994, 1997; Kleywegt & Jones, 1997), *e.g.* to flip a peptide plane, to replace a side chain by a rotamer conformation, to change side-chain torsion angles in order to optimize the fit to the density, to move groups of atoms, to use real-space refinement on a single residue or a zone of residues, to 'mutate' a residue to alanine *etc*. Together they constitute a toolbox with which many problems, once recognized, can be fixed relatively effortlessly (Kleywegt & Jones, 1997).

21.1.5. Preventing errors

As with everything else, when it comes to building a model of a protein, prevention of errors is the best medicine. Some general guidelines can be given (Dodson *et al.*, 1996; Kleywegt & Jones, 1997).

(1) Try to obtain the best possible set of data and the best possible set of phases for those data. If the structure has noncrystallographic symmetry (or if multiple crystal forms are available), use electron-density averaging to remove model bias and to reduce phase errors (Kleywegt & Read, 1997). In the absence of noncrystallographic symmetry, use maps that are biased by the model as little as possible [*e.g.* σ_A-weighted (Read, 1986) or omit maps (Bhat & Cohen, 1984; Bhat, 1988; Hodel *et al.*, 1992)]. If experimental phase information is available, keep and consult the experimental map(s). Experimental phases can also be used throughout the refinement process to alleviate or prevent some problems.

(2) Use databases to construct the initial model (or new parts of the model; Jones *et al.*, 1991; Kleywegt & Jones, 1998). All the crystallographer needs to do is to roughly place the C^α atoms in the density. The model-building program can then 'recycle' well refined high-resolution structures to place the main-chain atoms. Similarly, side-chain conformations should initially be chosen from the set of preferred rotamers for each residue type, perhaps in combination with a rigid-body rotation of the entire residue around its C^α atom and/or with minor adjustment of the torsion angles of long side chains (arginine, lysine *etc.*).

(3) After every cycle of refinement, carry out a critical analysis of the quality of the current model. This entails the calculation of properties such as those discussed in Section 21.1.3 and the inspection of the residues that are outliers for any of them, as described in Section 21.1.4. Be conservative during rebuilding, especially when the model is incomplete and possibly full of errors.

(4) Design a refinement protocol that is appropriate for the available data. If NCS restraints do not give a significantly better free *R* value than NCS constraints, then use constraints. If NCS restraints are to be employed, then use the experimental map to design a suitable NCS-restraint scheme (Kleywegt, 1999). Avoid the temptation to model alternative conformations in low-resolution maps or to place putative solvent molecules in every local maximum of the ($F_o - F_c$, α_c) difference map. In other words, be conservative and remember that the maxim 'where freedom is given, liberties are taken' is highly applicable to refinement programs (Hendrickson & Konnert, 1980; Kleywegt & Jones, 1995*b*).

(5) Adopt methodological advances as soon as they become available. Several innovations have only been slowly accepted by the mainstream (*e.g.* the use of databases in building and rebuilding, the use of the free *R* value, the use of electron-density averaging in molecular-replacement cases, bulk-solvent modelling). The most prominent recent development is the use of likelihood-based refinement programs (Bricogne & Irwin, 1996; Pannu & Read, 1996; Murshudov *et al.*, 1997; Adams *et al.*, 1997; Pannu *et al.*, 1998). These programs produce better models and maps and considerably reduce over-fitting (as assessed by the difference between the free and conventional *R* values).

(6) Most importantly, the crystallographer should be hypercritical towards the fruits of his or her own labour. Every intermediate model is a hypothesis to be shot down (Jones & Kjeldgaard, 1994). The crystallographer should be more critical than the supervisor, the supervisor more critical than the referee and the referee more critical than the casual reader. It goes without saying that the reader, casual or not, should have access to model coordinates, experimental data and electron-density maps!

21.1.6. Final model

Once the refinement is finished [*i.e.* once the ($F_o - F_c$, α_c) difference map is featureless (Cruickshank, 1950) and parameter shifts in further refinement cycles are negligibly small], three tasks remain: validation of the final model, description and analysis of the structure, and deposition of the model coordinates and the crystallographic data with the Protein Data Bank (Bernstein *et al.*, 1977).

Until a few years ago, validation of the final model typically entailed calculating the conventional *R* value, r.m.s. deviations from ideal values of bond lengths and angles, average temperature factors, and a Luzzati-type estimate of coordinate error. Kleywegt & Jones (1995*b*) showed that these statistics are not necessarily even remotely related to the actual quality of a model. Based on these criteria, a backwards-traced protein model was of higher apparent quality than a carefully refined correct model. After this, the realisation sunk in that the best validation criteria are those that assess aspects of the model that are 'orthogonal' to the information used during model refinement and rebuilding. For instance, the main-chain φ and ψ torsion angles are usually not restrained during refinement; this makes the Ramachandran plot such a powerful validation tool (Kleywegt & Jones, 1996*b*, 1998). Other examples of useful independent tests include the profile method of Eisenberg and co-workers (Lüthy *et al.*, 1992), the directional atomic contact analysis method of Vriend & Sander (1993) and the threading-potential method of Sippl (1993).

In general, all quality checks provide necessary, but in themselves insufficient, indications as to whether or not a model is essentially correct. A truly good model should make sense with respect to what is currently known about physics, chemistry, crystallography, protein structures, statistics and (last, but not least) biology and biochemistry (Kleywegt & Jones, 1995*a*). A good model will typically score well on most if not all validation criteria, whereas a poor one will score poorly on many criteria. The same is true at the level of residues: a poor or erroneous region in a model will be characterized by violations of many residue-level quality criteria (Kleywegt & Jones, 1997).

21.1.7. A compendium of quality criteria

In this section, some of the quality and validation criteria that have been used by macromolecular crystallographers are summarized (for more detailed information, the reader is referred to the primary literature). When judging how useful or powerful these criteria are in a certain case, one should keep in mind that any criterion that has been used explicitly or implicitly during model refinement (*e.g.* geometric restraints) or rebuilding (*e.g.* rotamer libraries) does *not* provide a truly independent check on the quality of the model.

Many, but not all, of the criteria discussed below pertain specifically to protein models. Comparatively little work has been performed on the validation of nucleic acid models, although there are indications that there is a need for such procedures (Schultze & Feigon, 1997). The situation would appear to be even worse for hetero-entities (*e.g.* ligands, inhibitors, cofactors, covalent attachments, saccharides, metals, ions; van Aalten *et al.*, 1996; Kleywegt & Jones, 1998).

21.1.7.1. Data quality

Although many quality and validation criteria have been developed for assessing coordinate sets of protein models, comparatively few criteria are available for assessing the quality of the crystallographic data.

21.1.7.1.1. Merging R values

Possibly the most common mistake in papers describing protein crystal structures is an incorrectly quoted formula for the merging R value (calculated during data reduction),

$$R_{\rm merge} = \sum_h \sum_i |I_{h,i} - \langle I_h\rangle| / \sum_h \sum_i I_{h,i},$$

where the outer sum (h) is over the unique reflections (in most implementations, only those reflections that have been measured more than once are included in the summations) and the inner sum (i) is over the set of independent observations of each unique reflection (Drenth, 1994). This statistic is supposed to reflect the spread of multiple observations of the intensity of the unique reflections (where the multiple observations may derive from symmetry-related reflections, different images or different crystals). Unfortunately, $R_{\rm merge}$ is a very poor statistic, since its value increases with increasing redundancy (Weiss & Hilgenfeld, 1997; Diederichs & Karplus, 1997), even though the signal-to-noise ratio of the average intensities will be higher as more observations are included (in theory, an N-fold increase of the number of independent observations should improve the signal-to-noise ratio by a factor of $N^{1/2}$). At high redundancy, the value of $R_{\rm merge}$ is directly related to the average signal-to-noise ratio (Weiss & Hilgenfeld, 1997): $R_{\rm merge} \simeq 0.8/\langle I/\sigma(I)\rangle$.

Diederichs & Karplus (1997) have suggested a number of alternative measures that lack most of the drawbacks of $R_{\rm merge}$. Their statistic $R_{\rm meas}$ is similar to $R_{\rm merge}$, but includes a correction for redundancy (m),

$$R_{\rm meas} = \sum_h [m/(m-1)]^{1/2} \sum_i |I_{h,i} - \langle I_h\rangle| / \sum_h \sum_i I_{h,i}.$$

Another statistic, the pooled coefficient of variation (PCV), is defined as

$${\rm PCV} = \sum_h \{[1/(m-1)] \sum_i (I_{h,i} - \langle I_h\rangle)^2\}^{1/2} / \sum_h \langle I_h\rangle.$$

Since PCV = $1/\langle I/\sigma(I)\rangle$, this quantity also provides an indication as to whether the standard deviations $\sigma(I)$ have been estimated appropriately. Finally, the statistic $R_{\rm mrgd\text{-}F}$, used for assessing the quality of the reduced data, enables a direct comparison of this merging R value with the refinement residuals R and $R_{\rm free}$.

Ideally, merging statistics should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.2. Completeness

Data completeness can be assessed by calculating what fraction of the unique reflections within a range of Bragg spacings that could in theory be observed has actually been measured. Ideally, completeness should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.3. Redundancy

Redundancy is defined as the number of independent observations (after merging of partial reflections) per unique reflection in the final merged and symmetry-reduced data set. Ideally, average redundancy should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.4. Signal strength

The average strength or significance of the observed intensities can be expressed in different ways. Values that are often quoted include the percentage of reflections for which $I/\sigma(I)$ exceeds a certain value (usually 3.0) and the average value of $I/\sigma(I)$. Ideally, these numbers should be quoted for all resolution shells (which should not be too broad), as well as for the entire data set. However, as a minimum, the values for the two extreme (low- and high-resolution) shells and for the entire data set should be reported.

21.1.7.1.5. Resolution

The nominal resolution limits of a data set are chosen by the crystallographer, usually at the data-processing stage, and ought to reflect the range of Bragg spacings for which useful intensity data have been collected. Unfortunately, owing to the subjective nature of this process, resolution limits cannot be compared meaningfully between data sets processed by different crystallographers. Careful crystallographers will take factors such as shell completeness, redundancy and $\langle I/\sigma(I)\rangle$ into account, whereas others may simply look up the minimum and maximum Bragg spacing of all observed reflections. Bart Hazes (personal communication) has suggested defining the effective resolution of a data set as that resolution at which the number of observed reflections would constitute a 100% complete data set. Alternatively, Vaguine *et al.* (1999) define the effective (or optical) resolution as the expected minimum distance between two resolved peaks in the electron-density map and calculate this quantity as $2\Delta_P/2^{1/2}$, where Δ_P is the width of the origin Patterson peak. One day, hopefully, the term 'resolution' will be replaced by an estimate of the information content of data sets. Randy Read (personal communication) has carried out preliminary work along these lines.

21.1.7.1.6. Unit-cell parameters

The accuracy of unit-cell parameters has been shown to be grossly overestimated for small-molecule crystal structures (Taylor & Kennard, 1986). Not intimidated by this observation, some macromolecular crystallographers routinely quote unit-cell axes of 100–200 Å with a precision of 0.01 Å. An analysis of several high-resolution protein crystal structures has revealed that surprisingly large errors in the unit-cell parameters appear to be quite common (at least if synchrotron sources are used for data collection; EU 3-D Validation Network, 1998). Such errors can be detected *a posteriori* by checking if the bond lengths in a model show any systematic, perhaps direction-dependent, variations from their target values.

21.1.7.1.7. Symmetry

From the symmetry of the diffraction pattern, the point-group symmetry of the crystal lattice can usually be derived. It is important to merge the data in the point group with the highest possible symmetry (usually assessed using merging statistics) in order to minimize the chance of making an incorrect space-group assignment (Marsh, 1995, 1997; Kleywegt *et al.*, 1996). Once the first data set has been processed, it is always useful to compute a self-rotation function. A non-origin peak of comparable strength to the origin peak will indicate that the true space group has higher symmetry. [Similarly, a self-Patterson function can be calculated at this stage to detect any purely translational NCS (Kleywegt & Read, 1997).] Once the final model is available, a search for possibly missed higher symmetry can be carried out, *e.g.* using the method developed by Hooft *et al.* (1994).

Sometimes crystallographic symmetry breaks down (pseudo-symmetry): an apparent higher symmetry at low resolution does not hold at higher resolution. In some cases, this is a consequence of the chemistry of the system studied (*e.g.* an asymmetric ligand bound by a symmetric protein dimer). In other cases, it may go undetected and complicate space-group determination and solution and refinement of the structure.

When it comes to space-group determination, many of the lessons learned by small-molecule crystallographers also apply to macromolecular crystallography (Marsh, 1995; Watkin, 1996).

21.1.7.2. Model quality, coordinates

Many criteria (and computer programs) are available to check for structural outliers based only on analysis of Cartesian coordinate sets.

21.1.7.2.1. Geometry and stereochemistry

The covalent geometry of a model can be assessed by comparing bond lengths and angles to a library of 'ideal' values. In the past, every refinement and modelling program had its own set of 'ideal' values. This even made it possible to detect (with 95% accuracy) with which program a model had been refined, simply by inspecting its covalent geometry (Laskowski, Moss & Thornton, 1993). Nowadays, standard sets of ideal bond lengths and bond angles derived from an analysis of small-molecule crystal structures from the CSD (Allen *et al.*, 1979) are available for proteins (Engh & Huber, 1991; Priestle, 1994) and nucleic acids (Parkinson *et al.*, 1996). For other entities, typical bond lengths and bond angles can be taken from tables of standard values (Allen *et al.*, 1987) or derived by other means (Kleywegt & Jones, 1998; Greaves *et al.*, 1999).

For bond lengths, the r.m.s. deviation from ideal values is invariably quoted. Deviations from ideality of bond angles can be expressed directly as an angular r.m.s. deviation or in terms of angle distances (*i.e.* the angle $\angle ABC$ is measured by the 1–3 distance $|AC|$; note that this distance is also implicitly dependent on the bond lengths $|AB|$ and $|BC|$). There are some indications that protein geometry cannot always be captured by assuming unimodal distributions (*i.e.* geometric features with only a single 'ideal' value). For example, Karplus (1996) found that the main-chain bond angle τ_3 ($\angle$N—C$^\alpha$—C) varies as a function of the main-chain torsion angles φ and ψ.

Chirality is another important criterion in the case of biomacromolecules: most amino-acid residues will have the L configuration for their C$^\alpha$ atom. Also, the C$^\beta$ atoms of threonine and isoleucine residues are chiral centres (IUPAC–IUB Commission on Biochemical Nomenclature, 1970; Morris *et al.*, 1992). Chirality can be assessed in terms of improper torsion angles or chiral volumes. For example, to check if the C$^\alpha$ atom of any residue other than glycine has the L configuration, the improper (or virtual) torsion angle C$^\alpha$—N—C—C$^\beta$ should have a value of about +34° (a value near −34° would indicate a D-amino acid). The torsion angle is called improper or virtual because it measures a torsion around something other than a covalent bond, in this case the N—C 'virtual bond'. The chiral volume is defined as the triple scalar product of the vectors from a central atom to three attached atoms (Hendrickson, 1985). For instance, the chiral volume of a C$^\alpha$ atom is defined as

$$V_{C^\alpha} = (\mathbf{r}_{\rm N} - \mathbf{r}_{C^\alpha}) \cdot [(\mathbf{r}_{\rm C} - \mathbf{r}_{C^\alpha}) \times (\mathbf{r}_{C^\beta} - \mathbf{r}_{C^\alpha})],$$

where $\mathbf{r}_X$ is the position vector of atom X. It should be noted that the chiral volume also implicitly depends on the bond lengths and angles involving the four atoms.

Another issue to consider is that of moieties that are necessarily planar (*e.g.* carboxylate groups, phenyl rings; Hooft *et al.*, 1996*a*). Again, planarity can be assessed in two different ways: by inspecting a set of (possibly improper) torsion angles and calculating their r.m.s. deviation from ideal values (*e.g.* all ring torsions in a perfectly flat phenyl ring should be 0°) or by fitting a least-squares plane through each set of atoms and calculating the r.m.s. distance of the atoms to that plane. Note that for double bonds, *cis* and *trans* configurations cannot be distinguished by deviations from a least-squares plane, but they can be distinguished by an appropriately defined torsion angle.

21.1.7.2.2. Torsion angles (dihedrals)

The conformation of the backbone of every non-terminal amino-acid residue is determined by three torsion angles, traditionally called φ (C_{i-1}—N_i—C_i^α—C_i), ψ (N_i—C_i^α—C_i—N_{i+1}) and ω (C_i^α—C_i—N_{i+1}—C_{i+1}^α). Owing to the peptide bond's partial double-bond character, the ω angle is restrained to values near 0° (*cis*-peptide) and 180° (*trans*-peptide). *Cis*-peptides are relatively rare and usually (but not always) occur if the next residue is a proline (Ramachandran & Mitra, 1976; Stewart *et al.*, 1990). The average ω-value for *trans*-peptides is slightly less than 180° (MacArthur & Thornton, 1996), but surprisingly large deviations have been observed in atomic resolution structures (Sevcik *et al.*, 1996; Merritt *et al.*, 1998). The ω angle therefore offers little in the way of validation checks, although values in the range of ±20 to ±160° should be treated with caution in anything but very high resolution models. The φ and ψ torsion angles, on the other hand, are much less restricted, but it has been known for a long time that owing to steric hindrance there are several clearly preferred combinations of φ, ψ values (Ramakrishnan & Ramachandran, 1965). This is true even for proline and glycine residues, although their distributions are atypical (Morris *et al.*, 1992). Also, an overwhelming majority of residues that are not in regular secondary-structure elements are found to have favourable φ, ψ torsion-angle combinations (Swindells *et al.*, 1995). For these reasons, the Ramachandran plot (essentially a φ, ψ scatter plot) is an extremely useful indicator of model quality (Weaver *et al.*, 1990; Laskowski, MacArthur *et al.*, 1993; MacArthur & Thornton, 1996; Kleywegt & Jones, 1996*b*; Kleywegt, 1996; Hooft *et al.*, 1997). Residues that have unusual φ, ψ torsion-angle combinations should be scrutinized by the crystallographer. If they have convincing electron density, there is probably a good structural or functional reason for the protein to tolerate the energetic strain that is associated with the unusual conformation (Herzberg & Moult, 1991). As a rule, the residue types that are most often found as outliers are serine, threonine, asparagine, aspartic acid and histidine (Gunasekaran *et al.*, 1996; Karplus, 1996). The quality of a model's Ramachandran plot is most convincingly illustrated by a figure. Alternatively, the fraction of residues in certain predefined areas of the plot (*e.g.* core regions) can be quoted, but in that case it is important to indicate which definition of such areas was used. Sometimes, one may also encounter a Balasubramanian plot, which is a linear φ, ψ plot as a function of the residue number (Balasubramanian, 1977).

In protein structures, the plane of the peptide bond can have two different orientations (approximately related by a 180° rotation around the virtual $C^{\alpha}-C^{\alpha}$ bond) that are both compatible with a *trans* configuration of the peptide (Jones *et al.*, 1991). The correct orientation can usually be deduced from the density of the carbonyl O atom or from the geometric requirements of regular secondary-structure elements (in α-helices, all carbonyl O atoms point towards the C-terminus of the helix; in β-strands, carbonyl O atoms usually alternate their direction). In other cases, *e.g.* in loops with poor density, the correct orientation may be more difficult to determine and errors are easily made. By comparing the local C^{α} conformation to a database of well refined high-resolution structures, unusual peptide orientations can be identified and, if required, corrected (through a 'peptide flip'; Jones *et al.*, 1991; Kleywegt & Jones, 1997, 1998). Since flipping the peptide plane between residues i and $i + 1$ changes the ψ angle of residue i and the φ angle of residue $i + 1$ by ~180°, erroneous peptide orientations may also lead to outliers in the Ramachandran plot (Kleywegt, 1996; Kleywegt & Jones, 1998).

All amino-acid residues whose side chain extends beyond the C^{β} atom contain one or more conformational side-chain torsion angles, termed χ_1 ($N-C^{\alpha}-C^{\beta}-X^{\gamma}$, where X may be carbon, sulfur or oxygen, depending on the residue type; if there are two γ atoms, the χ_1 torsion is calculated with reference to the atom with the lowest numerical identifier, *e.g.* $O^{\gamma 1}$ for threonine residues), χ_2 ($C^{\alpha}-C^{\beta}-X^{\gamma}-X^{\delta}$) *etc.* Early on, it was found that the values that these torsion angles assume in proteins are similar to those expected on the basis of simple energy calculations and that in addition certain combinations of χ_1, χ_2 values are clearly preferred (so-called rotamer conformations; Janin *et al.*, 1978; James & Sielecki, 1983; Ponder & Richards, 1987). Analogous to Ramachandran plots, χ_1, χ_2 scatter plots can be produced that show how well a protein's side-chain conformations conform to known preferences (Laskowski, MacArthur *et al.*, 1993; Carson *et al.*, 1994). Alternatively, a score can be computed for each residue that shows how similar its side-chain conformation is to that of the most similar rotamer for that residue type. This score can be calculated as an r.m.s. distance between corresponding side-chain atoms (Jones *et al.*, 1991; Zou & Mowbray, 1994; Kleywegt & Jones, 1998) or it can be expressed as an r.m.s. deviation of side-chain torsion-angle values from those of the most similar rotamer (Noble *et al.*, 1993).

Other torsion angles that have been used for validation purposes include the proline φ torsion (restricted to values near −65° owing to the geometry of the pyrrolidine ring; Morris *et al.*, 1992) and the χ_3 torsion in disulfide bridges (defined by the atoms $C^{\beta}-S-S'-C^{\beta'}$ and restricted to values near +95 and −85°; Morris *et al.*, 1992). In addition to the torsion-angle values of individual residues, pooled standard deviations of χ_1 and/or χ_2 torsions have been used for validation purposes (Morris *et al.*, 1992; Laskowski, MacArthur *et al.*, 1993).

To assess the 'geometric strain' in a model on a per-residue basis, the refinement program *X-PLOR* (Brünger, 1992*b*) can produce geometric pseudo-energy plots. In such a plot, the ratio of $E_{\rm geom}(i)/{\rm r.m.s.}(E_{\rm geom})$ is calculated as a function of the residue number i. The pseudo-energy term $E_{\rm geom}$ consists of the sums of the geometric and stereochemical pseudo-energy terms of the force field ($E_{\rm geom} = E_{\rm bonds} + E_{\rm angles} + E_{\rm dihedrals} + E_{\rm impropers}$), involving only the atoms of each residue.

It has been observed that the more high-resolution protein structures become available, the more 'well behaved' proteins turn out to be, *i.e.* the distributions of conformational torsion angles and torsion-angle combinations become even tighter than observed previously and the numerical averages tend to shift somewhat (Ponder & Richards, 1987; Kleywegt & Jones, 1998; EU 3-D Validation Network, 1998; MacArthur & Thornton, 1999; Walther & Cohen, 1999).

21.1.7.2.3. C^{α}-only models

Validation of C^{α}-only models may be necessary if such a model is retrieved from the PDB to be used in molecular replacement or homology modelling exercises; however, not many validation tools can handle such models (Kleywegt, 1997). The C^{α} backbone can be characterized by $C^{\alpha}-C^{\alpha}$ distances (~2.9 Å for a *cis*-peptide and ~3.8 Å for a *trans*-peptide), $C^{\alpha}-C^{\alpha}-C^{\alpha}$ pseudo-angles and $C^{\alpha}-C^{\alpha}-C^{\alpha}-C^{\alpha}$ pseudo-torsion angles (Kleywegt, 1997). The pseudo-angles and torsion angles turn out to assume certain preferred value combinations (Oldfield & Hubbard, 1994), much like the backbone φ and ψ torsions, and this can be employed for the validation of C^{α}-only models (Kleywegt, 1997). In addition to these straightforward methods, the mean-field approach of Sippl (1993) is also applicable to C^{α}-only models.

21.1.7.2.4. Contacts and environments

Hydrophobic, electrostatic and hydrogen-bonding interactions are the main stabilizing forces of protein structure. This leads to packing arrangements where hydrophobic residues tend to interact with each other, where charged residues tend to be involved in salt links and where hydrophilic residues prefer to interact with each other or to point out into the bulk solvent. Serious model errors will often lead to violations of such simple rules of thumb and introduce non-physical interactions (*e.g.* a charged arginine residue located inside a hydrophobic pocket; Kleywegt *et al.*, 1996) that serve as good indicators of model errors. Directional atomic contact analysis (Vriend & Sander, 1993) is a method in which these empirical notions have been formalized through database analysis. For every group of atoms in a protein, it yields a score which in essence expresses how 'comfortable' that group is in its environment in the model under scrutiny (compared with the expectations derived from the database). If a region in a model (or the entire model) has consistently low scores, this is a very strong indication of model errors. The *ERRAT* program is based on the same principle, but it is less specific in that it assesses only six types of non-bonded interactions (CC, CN, CO, NN, NO and OO; Colovos & Yeates, 1993).

Hydrogen-bonding analysis can often be used to determine the correct orientation of asparagine, glutamine and histidine residues (McDonald & Thornton, 1995). Similarly, an investigation of unsatisfied hydrogen-bonding potential can be used for validation purposes (Hooft *et al.*, 1996*b*), as can calculation of hydrogen-bonding energies (Morris *et al.*, 1992; Laskowski, MacArthur *et al.*, 1993).

Finally, a model should not contain unusually short non-bonded contacts. Although most refinement programs will restrain atoms from approaching one another too closely, if any serious violations remain they are worth investigating, since they may signal an underlying problem (*e.g.* erroneous omission of a disulfide restraint or incorrect side-chain assignment).

21.1.7.2.5. Noncrystallographic symmetry

Molecules that are related by noncrystallographic symmetry exist in very similar, but not identical, physical environments. This implies that their structures are expected to be quite similar, although different relative domain orientations and local variations may occur (*e.g.* owing to different crystal-packing interactions; Kleywegt, 1996). Many criteria have been developed to quantify the differences between (NCS) related models. Some, such as the r.m.s. distance (*e.g.* on all atoms, backbone atoms or C^{α} atoms) are based on distances between equivalent atoms, measured after a (to some extent arbitrary; Kleywegt, 1996) structural superpositioning operation has been performed. Others are based on a comparison of torsion angles, be it of main-chain φ, ψ angles [*e.g.* $\Delta\varphi$, $\Delta\psi$ plot (Korn & Rose, 1994); multiple-model Ramachandran plot (Kleywegt, 1996); $\sigma(\varphi)$, $\sigma(\psi)$ plot (Kleywegt, 1996); circular variance (Allen & Johnson, 1991) plots of φ and ψ (G. J. Kleywegt, unpublished results); Euclidian φ, ψ distances (Carson *et al.*, 1994) or pseudo-energy values (Carson *et al.*, 1994)] or side-chain χ_1, χ_2 angles [*e.g.* multiple-model χ_1, χ_2 plot (Kleywegt, 1996); $\sigma(\chi_1)$, $\sigma(\chi_2)$ plots (Kleywegt, 1996); circular variance (Allen & Johnson, 1991) plots of χ_1 and χ_2 (G. J. Kleywegt, unpublished results); Euclidian χ_1, χ_2 distances (Carson *et al.*, 1994) or pseudo-energy values (Carson *et al.*, 1994)]. Still other methods are based on analysing differences in contact-surface areas (Abagyan & Totrov, 1997), temperature factors (Kleywegt, 1996) or the geometry of the C^{α} backbone alone (Flocco & Mowbray, 1995; Kleywegt, 1996). Many of these methods can also be used to compare the structures of related molecules in different crystals or crystal forms (*e.g.* complexes, mutants).

21.1.7.2.6. Solvent molecules

Solvent molecules provide an excellent means of 'absorbing' problems in both the experimental data and the atomic model. Neither their position nor their temperature factor are usually restrained (other than by the data and restraints that prevent close contacts) and sometimes even their occupancy is refined. At a resolution of ~2 Å, crystallographers tend to model roughly one water molecule for every amino-acid residue and at 1.0 Å resolution this number increases to ~1.6 (Carugo & Bordo, 1999). When waters are placed, it should be ascertained that they can actually form hydrogen bonds, be it to protein atoms or to other water molecules. Considering that several ions that are isoelectronic with water (Na^+, NH_4^+) are often used in crystallization solutions, one should keep in mind the possibility that some entities that have been modelled as water molecules could be something else (Kleywegt & Jones, 1997). A method to check if water molecules could actually be sodium ions, based on the surrounding atoms, has been published (Nayal & Di Cera, 1996).

21.1.7.2.7. Miscellaneous

Many other coordinate-based methods for assessing the validity or correctness of protein models have been developed. These include the profile method of Eisenberg and co-workers (Bowie *et al.*, 1991; Lüthy *et al.*, 1992), the inspection of atomic volumes (Pontius *et al.*, 1996), and the use of threading and other potentials (Sippl, 1993; Melo & Feytmans, 1998; Maiorov & Abagyan, 1998). Some of these methods are described in more detail elsewhere in this volume. The program *WHAT IF* (Vriend, 1990) contains a large array of quality checks, many of which are not available in other programs, that span the spectrum from administrative checks to global quality indicators (Hooft *et al.*, 1996). During the refinement process, coordinate shifts can be used as a rough indication of 'quality' or, rather, convergence (Carson *et al.*, 1994; Kleywegt & Jones, 1996*a*). Crude models tend to undergo much larger changes during refinement than models that are essentially correct and complete. Also at the residue level, large coordinate shifts indicate residues that are worth a closer look.

Laskowski *et al.* (1994) have formulated single-number geometrical quality criteria, which they dubbed '*G* factors' in analogy to crystallographic *R* values. These *G* factors combine the results of a number of quality checks (covalent geometry, main-chain and side-chain torsion angles *etc.*) in a single number.

21.1.7.3. Model quality, temperature factors

In crystallographic refinement, atomic displacement parameters (ADPs; often referred to as temperature factors or *B* factors) model the effects of static and dynamic disorder. Except at high resolution (typically better than 1.5 Å), where there are sufficient observations to warrant refinement of anisotropic temperature factors, ADPs are usually constrained to be isotropic. The isotropic temperature factor *B* of an atom is related to the atom's mean-square displacement $\langle \Delta r^2 \rangle$ according to $B = 8\pi^2 \langle \Delta r^2 \rangle / 3$. Compared with the atomic coordinates, there are usually comparatively few restraints on temperature factors during refinement. Therefore, particularly at low resolution, temperature factors often function as 'error sinks' (Read, 1990). They absorb not only the effects of static and dynamic disorder, but also of various kinds of model errors.

Compared with the wealth of statistics that can be used to check and validate coordinates, there are relatively few methods available to assess how reasonable a model's temperature factors are. One obvious check is to see how well the average temperature factor of the model matches the value calculated from the data, using either a Wilson plot (Wilson, 1949) or the Patterson origin peak (Vaguine *et al.*, 1999). Since the average temperature factor of a model is usually not restrained, this is a useful check that has been used on several occasions to justify high average *B* factors. One should keep in mind that a low average *B* factor, *per se*, is not necessarily an indication of high model quality. For instance, a backwards-traced protein model can have a considerably lower average *B* factor than a correct model at a similar resolution (Kleywegt & Jones, 1995*b*). Average (and minimum and maximum) temperature-factor values can also be listed separately for various groups of atoms (*e.g.* individual protein or nucleic acid molecules, ligands, solvent molecules). A simple plot of residue-averaged temperature factors as a function of residue number may reveal regions of the molecule that have consistently high *B* factors, which may be a consequence of problems in the model (Kleywegt *et al.*, 1996).

Other statistics pertain to the r.m.s. differences in *B* factors between atoms that are somehow related, for example through a chemical bond (r.m.s. ΔB_{bonded}), through a 1–3 interaction or through noncrystallographic symmetry (possibly after correcting for any differences between the average B factors of the NCS-related molecules). Sometimes these statistics are calculated separately for main-chain and side-chain atoms. If the B factors of such related atoms have been restrained to be similar during refinement, these checks do not provide a convincing indication of the quality of the model. On the other hand, the B factors of atoms that have non-bonded interactions are usually not restrained to be similar, which renders the r.m.s. B-factor difference between such atoms (r.m.s. $\Delta B_{\text{non-bonded}}$) slightly more informative.

Since proteins tend to consist of a tightly packed core with more flexible regions at the surface, a radial B-factor plot (*i.e.* a plot of the average B factor of all atoms in a certain distance range from the centre of the molecule as a function of the distance) is expected to be shaped roughly like a half-parabola. Kuriyan & Weis (1991) used a ten-parameter isotropic rigid-molecule model of the mean-square atomic displacement (Schomaker & Trueblood, 1968). After obtaining values for the ten parameters (either by refinement against the structure-factor data or by fitting to the refined B factors of the model), the B factor of any atom can be calculated and depends only on its coordinates. They found that regions with large discrepancies between the refined and fitted B factors tend to be associated with errors or problems in a model.

Validation of anisotropic ADPs (Merritt, 1999), non-unit occupancies and H atoms, all of which are usually associated with high-resolution data, is still in its infancy. The validity of modelling anisotropic ADPs can be assessed by comparing the reduction of the conventional and free R values. If occupancies are used for multiple conformations of, for example, a side chain, the sum of the occupancies should be unity.

21.1.7.4. Model versus experimental data

21.1.7.4.1. R values

The traditional statistic used to assess how well a model fits the experimental data is the crystallographic R value,

$$R = \sum w \big| |F_o| - k|F_c| \big| / \sum |F_o|.$$

This statistic is closely related to the standard least-squares crystallographic residual $\sum w(|F_o| - k|F_c|)^2$ and its value can be reduced essentially arbitrarily by increasing the number of parameters used to describe the model (*e.g.* by refining anisotropic ADPs and occupancies for all atoms) or, conversely, by reducing the number of experimental observations (*e.g.* through resolution and σ cutoffs) or the number of restraints imposed on the model. Therefore, the conventional R value is only meaningful if the number of experimental observations and restraints greatly exceeds the number of model parameters. In 1992, Brünger introduced the free R value (R_{free}; Brünger, 1992*a*, 1993, 1997; Kleywegt & Brünger, 1996), whose definition is identical to that of the conventional R value, except that the free R value is calculated for a small subset of reflections that are not used in the refinement of the model. The free R value, therefore, measures how well the model predicts experimental observations that are not used to fit the model (cross-validation). Until a few years ago, a conventional R value below 0.25 was generally considered to be a sign that a model was essentially correct (Brändén & Jones, 1990). While this is probably true at high resolution, it was subsequently shown for several intentionally mistraced models that these can be refined to deceptively low conventional R values (Jones *et al.*, 1991; Kleywegt & Jones, 1995*b*; Kleywegt & Brünger, 1996). Brünger suggests a threshold value of 0.40 for the free R value, *i.e.* models with free R values greater than 0.40 should be treated with caution (Brünger, 1997). Tickle and coworkers have developed methods to estimate the expected value of R_{free} in least-squares refinement (Tickle *et al.*, 1998). Since the difference between the conventional and free R value is partly a measure of the extent to which the model over-fits the data (*i.e.* some aspects of the model improve the conventional but not the free R value and are therefore likely to fit noise rather than signal in the data), this difference $R_{\text{free}} - R$ should be small (Kleywegt & Jones, 1995*a*; Kleywegt & Brünger, 1996). Alternatively, the R_{free} ratio (defined as R_{free}/R; Tickle *et al.*, 1998) should be close to unity. Various practical aspects of the use of the free R value have been discussed by Kleywegt & Brünger (1996) and by Brünger (1997).

Self-validation is an alternative to cross-validation and in the case of crystallographic refinement, the Hamilton test (Hamilton, 1965) is a prime example of this. This method enables one to assess whether a reduction in the R value is statistically significant given the increase in the number of degrees of freedom. Application of this test in the case of macromolecules is compounded by the difficulty of estimating the effect of the combined set of restraints on the (effective) number of degrees of freedom, but some information can nevertheless be gained from such an analysis (Bacchi *et al.*, 1996).

21.1.7.4.2. Real-space fits

The fit of a model to the data can also be assessed in real space, which has the advantage that it can be performed for arbitrary sets of atoms (*e.g.* for every residue separately). Jones *et al.* (1991) introduced the real-space R value, which measures the similarity of a map calculated directly from the model (ρ_c) and one which incorporates experimental data (ρ_o) as

$$R = \sum |\rho_o - \rho_c| / \sum |\rho_o + \rho_c|,$$

where the sums extend over all grid points in the map that surround the selected set of atoms. The real-space fit can also be expressed as a correlation coefficient (Jones & Kjeldgaard, 1997), which has the advantage that no scaling of the two densities is necessary. Chapman (1995) described a modification in which the density calculated from the model is derived by Fourier transformation of resolution-truncated atomic scattering factors.

The program *SFCHECK* (Vaguine *et al.*, 1999) implements several variations on the real-space fit. The normalized average displacement measures the tendency of groups of atoms to move away from their current position. The density correlation is a modification of the real-space correlation coefficient. The residue-density index is calculated as the geometric mean of the density values of a set of atoms, divided by the average density of all atoms in the model. It therefore measures how high the electron-density level is for the set of atoms considered (*e.g.* all side-chain atoms of a residue). The connectivity index is identical to the residue-density index, but is calculated only for the N, C^{α} and C atoms. It thus provides an indication of the continuity of the main-chain electron density.

21.1.7.4.3. Coordinate error estimates

Since a measurement without an error estimate is not a measurement, crystallographers are keen to assess the estimated errors in the atomic coordinates and, by extension, in the atomic positions, bond lengths *etc.* In principle, upon convergence of a least-squares refinement, the variances and covariances of the model parameters (coordinates, ADPs and occupancies) may be obtained through inversion of the least-squares full matrix

(Sheldrick, 1996; Ten Eyck, 1996; Cruickshank, 1999). In practice, however, this is seldom performed as the matrix inversion requires enormous computational resources. Therefore, one of a battery of (sometimes quasi-empirical) approximations is usually employed.

For a long time, the elegant method of Luzzati (1952) has been used for a different purpose (namely, to estimate average coordinate errors of macromolecular models) than that for which it was developed (namely, to estimate the positional changes required to reach a zero R value, using several assumptions that are not valid for macromolecules; Cruickshank, 1999). A Luzzati plot is a plot of R value *versus* $2\sin\theta/\lambda$, and a comparison with theoretical curves is used to estimate the average positional error. Considering the problems with conventional R values (discussed in Section 21.1.7.4.1), Kleywegt *et al.* (1994) instead plotted free R values to obtain a cross-validated error estimate. This intuitive modification turned out to yield fairly reasonable values in practice (Kleywegt & Brünger, 1996; Brünger, 1997). Read (1986, 1990) estimated coordinate error from σ_A plots; the cross-validated modification of this method also yields reasonable error estimates (Brünger, 1997).

Cruickshank, almost 50 years after his work on the precision of small-molecule crystal structures (Cruickshank, 1949), introduced the diffraction-component precision index (DPI; Dodson *et al.*, 1996; Cruickshank, 1999) to estimate the coordinate or positional error of an atom with a B factor equal to the average B factor of the whole structure. In several cases for which full-matrix error estimates are available, the DPI gives quantitatively similar results. *SFCHECK* (Vaguine *et al.*, 1999) calculates both the DPI and Cruickshank's 1949 statistic (now termed the 'expected maximal error') based on the slope and the curvature of the electron-density map.

21.1.7.4.4. Noncrystallographic symmetry

Despite the multitude of criteria for assessing conformational differences between related molecules, there was until recently no objective way to assess whether such differences were a true reflection of the experimental data or a manifestation of refinement artifacts (Kleywegt & Jones, 1995*b*; Kleywegt, 1996). However, it has been found that electron-density maps calculated with experimental phases (or, at least, phases that are biased as little as possible by the model) and amplitudes can be used to correlate expected similarities (based on the data) with observed ones (manifest in the final refined models; Kleywegt, 1999). This method uses a local density-correlation map, as introduced by Read (Vellieux *et al.*, 1995), to measure the local similarity of the density of two or more models on a per-atom or per-residue basis. By comparing these values to the observed structural differences in the final models, it is relatively easy to check if the latter differences are warranted by the information contained in the experimental data (Kleywegt, 1999).

21.1.7.4.5. Difference density quality

van den Akker & Hol (1999) described a method (called *DDQ*, standing for difference density quality) to assess the local and global quality of a model based on analysis of an $(F_o - F_c, \alpha_c)$ map calculated after omission of all water molecules. In this method, the map and model are used to calculate several scores. One score assesses the presence or absence of favourably positioned water peaks near polar and apolar atoms. Other scores provide a measure for the presence or absence of positive and negative shift peaks that may indicate incorrect coordinates, temperature factors or occupancies. The scores can be averaged per residue or for an entire model and can be used to detect problems in models. The method appears to be applicable to ~3 Å resolution.

21.1.7.5. Accountancy

The opinions as to what constitutes an error in a model vary somewhat in the community [compare Hooft *et al.* (1996) and Jones *et al.* (1996), for instance], but most people would agree that a crystallographic error is one that requires access to the experimental data for its verification, and whose correction alters the calculated structure factors (*e.g.* position, B factor, occupancy or scattering factors of one or more atoms). In addition to this, there are nomenclature rules and conventions to which a model that is made publicly available should adhere. Separate from this is the issue of the (more clerical) validation of public database entries ('PDB files'; Hooft *et al.*, 1994, 1996) which, while important to maintain the integrity of these databases, ultimately ought to be the responsibility of the database curators (Jones *et al.*, 1996; Keller *et al.*, 1998; Abola *et al.*, 2000).

21.1.8. Future

During the 1990s, the field of protein model validation matured rapidly (MacArthur *et al.*, 1994; EU 3-D Validation Network, 1998; Laskowski *et al.*, 1998) and further fundamental breakthroughs seem unlikely at present (although it would be highly desirable to be able to calculate and compare the information content of experimental data and models alike). In contrast, work on the validation of nucleic acid models (Schultze & Feigon, 1997) and hetero-entities (Kleywegt & Jones, 1998) has only just begun. In addition, there is still scope for further development of validation methods that use both the atomic model and the crystallographic data. In addition, the increasing number of structures that are solved at (near-)atomic resolution may lead to an adjustment of some validation criteria, *e.g.* of 'ideal' geometric target values, rotamer libraries *etc.* Also, validation of model aspects typically associated with very high resolution studies (refined occupancies, alternative conformations, anisotropic ADPs, H atoms) is still poorly developed. An increased understanding and appreciation of factors that determine model quality (and knowledge of how to measure them) will be important for the development of more automatic methods for protein structure determination. This in turn will enable 'black-box' high-throughput protein crystallography to become a reality, at least for 'run-of-the-mill' structures.

The author gratefully acknowledges the many useful discussions about validation with Alwyn Jones (Uppsala), Axel Brunger (Yale), Eleanor Dodson (York), Randy Read (Cambridge), Carl-Ivar Brändén (Stockholm), the members of the EU-funded 3-D Validation Network and many other colleagues in the field. This work was supported by the Swedish Foundation for Strategic Research (SSF), its Structural Biology Network (SBNet) and the EU-funded 3-D Validation Network.

References

Aalten, D. M. F. van, Bywater, R., Findlay, J. B. C., Hendlich, M., Hooft, R. W. W. & Vriend, G. (1996). *PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. J. Comput. Aided Mol. Des.* **10**, 255–262.

Abagyan, R. A. & Totrov, M. M. (1997). *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J. Mol. Biol.* **268**, 678–685.

Abola, E. E., Bairoch, A., Barker, W. C., Beck, S., Benson, D. A., Berman, H., Cantor, C., Doubet, S., Hubbard, T. J. P., Jones, T. A., Kleywegt, G. J., Kolaskar, A. S., van Kuik, A., Lesk, A. M., Mewes, H. W., Neuhaus, D., Pfeiffer, F., Ten Eyck, L. F., Simpson, R. J., Stoesser, G., Sussman, J. L., Tateno, Y., Tsugita, A., Ulrich, E. L. & Vliegenthart, J. F. G. (2000). *Quality control in databanks for molecular biology. BioEssays*, **22**, 1024–1034.

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1997). *Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Akker, F. van den & Hol, W. G. J. (1999). *Difference density quality (DDQ): a method to assess the global and local correctness of macromolecular crystal structures. Acta Cryst.* D**55**, 206–218.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. Acta Cryst.* B**35**, 2331–2339.

Allen, F. H. & Johnson, O. (1991). *Automated conformational analysis from crystallographic data. 4. Statistical descriptors for a distribution of torsion angles. Acta Cryst.* B**47**, 62–67.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Ammon, H. L., Weber, I. T., Wlodawer, A., Harrison, R. W., Gilliland, G. L., Murphy, K. C., Sjölin, L. & Roberts, J. (1988). *Preliminary crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. J. Biol. Chem.* **263**, 150–156.

Bacchi, A., Lamzin, V. S. & Wilson, K. S. (1996). *A self-validation technique for protein structure refinement: the extended Hamilton test. Acta Cryst.* D**52**, 641–646.

Balasubramanian, R. (1977). *New type of representation for mapping chain-folding in protein molecules. Nature (London)*, **266**, 856–857.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol.* **112**, 535–542.

Bhat, T. N. (1988). *Calculation of an OMIT map. J. Appl. Cryst.* **21**, 279–281.

Bhat, T. N. & Cohen, G. H. (1984). *OMITMAP: an electron density map suitable for the examination of errors in a macromolecular model. J. Appl. Cryst.* **17**, 244–248.

Borgstahl, G. E. O., Williams, D. R. & Getzoff, E. D. (1995). *1.4 Å structure of photoactive yellow protein, a cytosolic photoreceptor: unusual fold, active site, and chromophore. Biochemistry*, **34**, 6278–6287.

Bott, R. & Sarma, R. (1976). *Crystal structure of turkey egg-white lysozyme: results of the molecular replacement method at 5 Å resolution. J. Mol. Biol.* **106**, 1037–1046.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). *A method to identify protein sequences that fold into a known three-dimensional structure. Science*, **253**, 164–170.

Brändén, C.-I. & Jones, T. A. (1990). *Between objectivity and subjectivity. Nature (London)*, **343**, 687–689.

Bricogne, G. & Irwin, J. (1996). *Maximum-likelihood refinement of incomplete models with BUSTER + TNT.* In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992*a*). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR. A System for Crystallography and NMR.* Yale University, New Haven, Connecticut, USA.

Brünger, A. T. (1993). *Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Cryst.* D**49**, 24–36.

Brünger, A. T. (1997). *The free R value: a more objective statistic for crystallography. Methods Enzymol.* **277**, 366–396.

Carson, M., Buckner, T. W., Yang, Z., Narayana, S. V. L. & Bugg, C. E. (1994). *Error detection in crystallographic models. Acta Cryst.* D**50**, 900–909.

Carugo, O. & Bordo, D. (1999). *How many water molecules can be detected by protein crystallography? Acta Cryst.* D**55**, 479–483.

Chapman, M. S. (1995). *Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. Acta Cryst.* A**51**, 69–80.

Colovos, C. & Yeates, T. O. (1993). *Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci.* **2**, 1511–1519.

Cruickshank, D. W. J. (1949). *The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1950). *The convergence of the least-squares or Fourier refinement methods. Acta Cryst.* **3**, 10–13.

Cruickshank, D. W. J. (1999). *Remarks about protein structure precision. Acta Cryst.* D**55**, 583–601.

Diederichs, K. & Karplus, P. A. (1997). *Improved R-factors for diffraction data analysis in macromolecular crystallography. Nature Struct. Biol.* **4**, 269–275.

Dodson, E., Kleywegt, G. J. & Wilson, K. S. (1996). *Report of a workshop on the use of statistical validators in protein X-ray crystallography. Acta Cryst.* D**52**, 228–234.

Drenth, J. (1994). *Principles of Protein X-ray Crystallography.* New York: Springer–Verlag.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

EU 3-D Validation Network (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol.* **276**, 417–436.

Flocco, M. M. & Mowbray, S. L. (1995). *Cα-based torsion angles: a simple tool to analyze protein conformational changes. Protein Sci.* **4**, 2118–2122.

Greaves, R. B., Vagin, A. A. & Dodson, E. J. (1999). *Automated production of small-molecule dictionaries for use in crystallographic refinements. Acta Cryst.* D**55**, 1335–1339.

Gunasekaran, K., Ramakrishnan, C. & Balaram, P. (1996). *Disallowed Ramachandran conformations of amino acid residues in protein structures. J. Mol. Biol.* **264**, 191–198.

Hamilton, W. C. (1965). *Significance tests on the crystallographic R factor. Acta Cryst.* **18**, 502–510.

Hendrickson, W. A. (1985). *Stereochemically restrained refinement of macromolecular structures. Methods Enzymol.* **115**, 252–270.

Hendrickson, W. A. & Konnert, J. H. (1980). *Incorporation of stereochemical information into crystallographic refinement.* In *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 13.01–13.25. Bangalore: Indian Academy of Science.

Herzberg, O. & Moult, J. (1991). *Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins Struct. Funct. Genet.* **11**, 223–229.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Model bias in macromolecular crystal structures. Acta Cryst.* A**48**, 851–858.

Hoier, H., Schlömann, M., Hammer, A., Glusker, J. P., Carrell, H. L., Goldman, A., Stezowski, J. J. & Heinemann, U. (1994). *Crystal structure of chloromuconate cycloisomerase from Alcaligenes eutrophus JMP134 (pJP4) at 3 Å resolution. Acta Cryst.* D**50**, 75–84.

Hooft, R. W. W., Sander, C. & Vriend, G. (1994). *Reconstruction of symmetry-related molecules from Protein Data Bank (PDB) files. J. Appl. Cryst.* **27**, 1006–1009.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996*a*). *Verification of protein structures: side-chain planarity. J. Appl. Cryst.* **29**, 714–716.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996*b*). *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins Struct. Funct. Genet.* **26**, 363–376.

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci.* **13**, 425–430.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures. Nature (London)*, **381**, 272.

IUPAC–IUB Commission on Biochemical Nomenclature (1970). *Abbreviations and symbols for the description of the conformation of polypeptide chains. J. Mol. Biol.* **52**, 1–17.

James, M. N. G. & Sielecki, A. R. (1983). *Structure and refinement of penicillopepsin at 1.8 Å resolution. J. Mol. Biol.* **163**, 299–361.

Janin, J. (1990). *Errors in three dimensions. Biochimie*, **72**, 705–709.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). *Conformation of amino acid side-chains in proteins. J. Mol. Biol.* **125**, 357–386.

Jones, T. A. & Kjeldgaard, M. (1994). *Making the first trace with O.* In *Proceedings of the CCP4 Study Weekend. From First Map to Final*

Model, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 1–13. Warrington: Daresbury Laboratory.

Jones, T. A. & Kjeldgaard, M. (1997). *Electron density map interpretation. Methods Enzymol.* **277**, 173–208.

Jones, T. A., Kleywegt, G. J. & Brünger, A. T. (1996). *Storing diffraction data. Nature (London)*, **381**, 18–19.

Jones, T. A. & Thirup, S. (1986). *Using known substructures in protein model building and crystallography. EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Karplus, P. A. (1996). *Experimentally observed conformation-dependent geometry and hidden strain in proteins. Protein Sci.* **5**, 1406–1420.

Keller, P. A., Henrick, K., McNeil, P., Moodie, S. & Barton, G. J. (1998). *Deposition of macromolecular structures. Acta Cryst.* D**54**, 1105–1108.

Kelly, J. A., Dideberg, O., Charlier, P., Wery, J. P., Libert, M., Moews, P. C., Knox, J. R., Duez, C., Fraipoint, C., Joris, B., Dusart, J., Frère, J. M. & Ghuysen, J. M. (1986). *On the origin of bacterial resistance to penicillin: comparison of a β-lactamase and a penicillin target. Science*, **231**, 1429–1431.

Kelly, J. A. & Kuzin, A. P. (1995). *The refined crystallographic structure of a DD-peptidase penicillin-target enzyme at 1.6 Å resolution. J. Mol. Biol.* **254**, 223–236.

Kleywegt, G. J. (1996). *Use of non-crystallographic symmetry in protein structure refinement. Acta Cryst.* D**52**, 842–857.

Kleywegt, G. J. (1997). *Validation of protein models from Cα coordinates alone. J. Mol. Biol.* **273**, 371–376.

Kleywegt, G. J. (1999). *Experimental assessment of differences between related protein crystal structures. Acta Cryst.* D**55**, 1878–1884.

Kleywegt, G. J., Bergfors, T., Senn, H., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Crystal structures of cellular retinoic acid binding proteins I and II in complex with all-trans-retinoic acid and a synthetic retinoid. Structure*, **2**, 1241–1258.

Kleywegt, G. J. & Brünger, A. T. (1996). *Checking your imagination: applications of the free R value. Structure*, **4**, 897–904.

Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *A re-evaluation of the crystal structure of chloromuconate cycloisomerase. Acta Cryst.* D**52**, 858–863.

Kleywegt, G. J. & Jones, T. A. (1995*a*). *Braille for pugilists.* In *Proceedings of the CCP4 Study Weekend. Making the Most of Your Model*, edited by W. N. Hunter, J. M. Thornton & S. Bailey, pp. 11–24. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1995*b*). *Where freedom is given, liberties are taken. Structure*, **3**, 535–540.

Kleywegt, G. J. & Jones, T. A. (1996*a*). *Efficient rebuilding of protein structures. Acta Cryst.* D**52**, 829–832.

Kleywegt, G. J. & Jones, T. A. (1996*b*). *Phi/Psi-chology: Ramachandran revisited. Structure*, **4**, 1395–1400.

Kleywegt, G. J. & Jones, T. A. (1997). *Model-building and refinement practice. Methods Enzymol.* **277**, 208–230.

Kleywegt, G. J. & Jones, T. A. (1998). *Databases in protein crystallography. Acta Cryst.* D**54**, 1119–1131.

Kleywegt, G. J. & Read, R. J. (1997). *Not your average density. Structure*, **5**, 1557–1569.

Kleywegt, G. J., Zou, J. Y., Divne, C., Davies, G. J., Sinning, I., Ståhlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T. T. & Jones, T. A. (1997). *The crystal structure of the catalytic core domain of endoglucanase I from Trichoderma reesei at 3.6 Å resolution, and a comparison with related enzymes. J. Mol. Biol.* **272**, 383–397.

Korn, A. P. & Rose, D. R. (1994). *Torsion angle differences as a means of pinpointing local polypeptide chain trajectory changes for identical proteins in different conformational states. Protein Eng.* **7**, 961–967.

Kuriyan, J. & Weis, W. I. (1991). *Rigid protein motion as a model for crystallographic temperature factors. Proc. Natl Acad. Sci. USA*, **88**, 2773–2777.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1994). *Evaluation of protein coordinate data sets.* In *Proceedings of the CCP4 Study Weekend. From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 149–159. Warrington: Daresbury Laboratory.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1998). *Validation of protein models derived from experiment. Curr. Opin. Struct. Biol.* **8**, 631–639.

Laskowski, R. A., Moss, D. S. & Thornton, J. M. (1993). *Main-chain bond lengths and bond angles in protein structures. J. Mol. Biol.* **231**, 1049–1067.

Lubkowski, J., Wlodawer, A., Housset, D., Weber, I. T., Ammon, H. L., Murphy, K. C. & Swain, A. L. (1994). *Refined crystal structure of Acinetobacter glutaminasificans glutaminase-asparaginase. Acta Cryst.* D**50**, 826–832.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). *Assessment of protein models with three-dimensional profiles. Nature (London)*, **356**, 83–85.

Luzzati, V. (1952). *Traitement statistique des erreurs dans la determination des structures crystallines. Acta Cryst.* **5**, 802–810.

MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). *Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy. Curr. Opin. Struct. Biol.* **4**, 731–737.

MacArthur, M. W. & Thornton, J. M. (1996). *Deviations from planarity of the peptide bond in peptides and proteins. J. Mol. Biol.* **264**, 1180–1195.

MacArthur, M. W. & Thornton, J. M. (1999). *Protein side-chain conformation: a systematic variation of χ_1 mean values with resolution – a consequence of multiple rotameric states? Acta Cryst.* D**55**, 994–1004.

McDonald, I. K. & Thornton, J. M. (1995). *The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng.* **8**, 217–224.

McRee, D. E., Tainer, J. A., Meyer, T. E., van Beeumen, J., Cusanovich, M. A. & Getzoff, E. D. (1989). *Crystallographic structure of a photoreceptor protein at 2.4 Å resolution. Proc. Natl Acad. Sci. USA*, **86**, 6533–6537.

Maiorov, V. & Abagyan, R. (1998). *Energy strain in three-dimensional protein structures. Fold. Des.* **3**, 259–269.

Marsh, R. E. (1995). *Some thoughts on choosing the correct space group. Acta Cryst.* B**51**, 897–907.

Marsh, R. E. (1997). *The perils of Cc revisited. Acta Cryst.* B**53**, 317–322.

Melo, F. & Feytmans, E. (1998). *Assessing protein structures with a non-local atomic interaction energy. J. Mol. Biol.* **277**, 1141–1152.

Merritt, E. A. (1999). *Expanding the model: anisotropic displacement parameters in protein structure refinement. Acta Cryst.* D**55**, 1109–1117.

Merritt, E. A., Kuhn, P., Sarfaty, S., Erbe, J. L., Holmes, R. K. & Hol, W. G. J. (1998). *The 1.25 Å resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. J. Mol. Biol.* **282**, 1043–1059.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Stereochemical quality of protein structure coordinates. Proteins Struct. Funct. Genet.* **12**, 345–364.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Nayal, M. & Di Cera, E. (1996). *Valence screening of water in protein crystals reveals potential Na$^+$ binding sites. J. Mol. Biol.* **256**, 228–234.

Noble, M. E. M., Zeelen, J. P., Wierenga, R. K., Mainfroid, V., Goraj, K., Gohimont, A. C. & Martial, J. A. (1993). *Structure of triosephosphate isomerase from Escherichia coli determined at 2.6 Å resolution. Acta Cryst.* D**49**, 403–417.

Nunn, R. S., Artymiuk, P. J., Baker, P. J., Rice, D. W. & Hunter, C. N. (1995). *Retraction – Purification and crystallization of the light harvesting LH1 complex from Rhodobacter sphaeroides. J. Mol. Biol.* **252**, 153.

Oldfield, T. J. & Hubbard, R. E. (1994). *Analysis of Cα geometry in protein structures. Proteins Struct. Funct. Genet.* **18**, 324–337.

Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Incorporation of prior phase information strengthens maximum-likelihood structure refinement. Acta Cryst.* D**54**, 1285–1294.

Pannu, N. S. & Read, R. J. (1996). *Improved structure refinement through maximum likelihood. Acta Cryst.* A**52**, 659–668.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.

Ponder, J. W. & Richards, F. M. (1987). *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol.* **193**, 775–791.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol.* **264**, 121–136.

Priestle, J. P. (1994). *Stereochemical dictionaries for protein structure refinement and model building. Structure*, **2**, 911–913.

Ramachandran, G. N. & Mitra, A. K. (1976). *An explanation for the rare occurrence of cis peptide units in proteins and polypeptides. J. Mol. Biol.* **107**, 85–92.

Ramachandran, G. N. & Srinivasan, R. (1961). *An apparent paradox in crystal structure analysis. Nature (London)*, **190**, 159–161.

Ramakrishnan, C. & Ramachandran, G. N. (1965). *Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. Biophys. J.* **5**, 909–933.

Read, R. J. (1986). *Improved Fourier coefficients for maps using phases from partial structures with errors. Acta Cryst.* A**42**, 140–149.

Read, R. J. (1990). *Structure-factor probabilities for related structures. Acta Cryst.* A**46**, 900–912.

Read, R. J. (1994). *Model bias and phase combination.* In *Proceedings of the CCP4 Study Weekend. From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 31–40. Warrington: Daresbury Laboratory.

Read, R. J. (1997). *Model phases: probabilities and bias. Methods Enzymol.* **277**, 110–128.

Schomaker, V. & Trueblood, K. N. (1968). *On the rigid-body motion of molecules in crystals. Acta Cryst.* B**24**, 63–76.

Schultze, P. & Feigon, R. (1997). *Chirality errors in nucleic acid structures. Nature (London)*, **387**, 668.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Ribonuclease from Streptomyces aureofaciens at atomic resolution. Acta Cryst.* D**52**, 327–344.

Sheldrick, G. M. (1996). *Least-squares refinement of macromolecules: estimated standard deviations, NCS restraints and factors affecting convergence.* In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 47–58. Warrington: Daresbury Laboratory.

Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., Board, P. G., Olin, B., Mannervik, B. & Jones, T. A. (1993). *Structure determination and refinement of human alpha class glutathione transferase A1-1, and a comparison with the mu and pi class enzymes. J. Mol. Biol.* **232**, 192–212.

Sippl, M. J. (1993). *Recognition of errors in three-dimensional structures of proteins. Proteins Struct. Funct. Genet.* **17**, 355–362.

Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *Occurrence and role of cis peptide bonds in protein structures. J. Mol. Biol.* **214**, 253–260.

Swindells, M. B., MacArthur, M. W. & Thornton, J. M. (1995). *Intrinsic φ, ψ propensities of amino acids, derived from the coil regions of known structures. Nature Struct. Biol.* **2**, 596–603.

Taylor, R. & Kennard, O. (1986). *Accuracy of crystal structure error estimates. Acta Cryst.* B**42**, 112–120.

Ten Eyck, L. F. (1996). *Full matrix least squares.* In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 37–45. Warrington: Daresbury Laboratory.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. Acta Cryst.* D**54**, 547–557.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst.* D**55**, 191–205.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *DEMON/ANGEL: a suite of programs to carry out density modification. J. Appl. Cryst.* **28**, 347–351.

Vriend, G. (1990). *WHAT IF: a molecular modeling and drug design program. J. Mol. Graphics*, **8**, 52–56.

Vriend, G. & Sander, C. (1993). *Quality control of protein models: directional atomic contact analysis. J. Appl. Cryst.* **26**, 47–60.

Walther, D. & Cohen, F. E. (1999). *Conformational attractors on the Ramachandran map. Acta Cryst.* D**55**, 506–517.

Watkin, D. (1996). *Pseudo symmetry.* In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 171–184. Warrington: Daresbury Laboratory.

Weaver, L. H., Tronrud, D. E., Nicholson, H. & Matthews, B. W. (1990). *Some uses of the Ramachandran (φ, ψ) diagram in the structural analysis of lysozymes. Curr. Sci.* **59**, 833–837.

Weiss, M. S. & Hilgenfeld, R. (1997). *On the use of the merging R factor as a quality indicator for X-ray data. J. Appl. Cryst.* **30**, 203–205.

Wilson, A. J. C. (1949). *The probability distribution of X-ray intensities. Acta Cryst.* **2**, 318–321.

Zou, J. Y. & Mowbray, S. L. (1994). *An evaluation of the use of databases in protein structure refinement. Acta Cryst.* D**50**, 237–249.

International Tables for Crystallography (2012). Vol. F, Chapter 21.2, pp. 662–676.

Chapter 21.2. Assessing the quality of macromolecular structures

S. J. Wodak, A. A. Vagin, J. Richelle, U. Das, J. Pontius and H. M. Berman

21.2.1. Introduction

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two major techniques that provide detailed information on the atomic structure of macromolecules. Usually, however, the data obtained from these techniques are not of high enough resolution to define the atomic positions of a macromolecule with sufficient precision. Deriving the atomic models from the experimental data therefore involves sophisticated optimization (refinement) procedures, in which constraints based on prior knowledge about the chemical structure of the molecule and its conformational properties are applied. The resulting models are therefore prone to errors, which fall into two broad categories: systematic errors caused by biases during the structure determination and refinement procedures, and random errors which affect the precision of the model. Moreover, the quality of the model can vary in different regions of the structure, often due to higher local conformational or thermal disorder in certain parts.

With the rapid growth in the number of structures of macromolecules determined and the spreading use of structural information in different areas of science, the availability of objective criteria and methods for evaluating the quality of these structures has become a very important requirement.

A variety of validation procedures have been proposed by many groups (for recent reviews, see MacArthur *et al.*, 1994, and Laskowski *et al.*, 1998). The procedures involve two main approaches. One approach comprises procedures that validate the geometric and conformational parameters of the final model. This is done by measuring the extent to which the parameters deviate from standard values, derived from crystals of small molecules or from a set of high-quality structures of other macromolecules. The main limitation of this approach is that the quality of a model is defined by comparison with other known models without taking into account the experimental data. This harbours the danger of considering unusual conformations related to biological function as errors in the model or of accepting as 'normal' only what has already been observed before. The second and most important approach by far comprises procedures that take into account the experimental data and evaluate the agreement of the atomic model with these data. These procedures can, in principle, evaluate systematic errors and biases that affect the global quality of the model and can also detect local imprecision. The most commonly cited measures of agreement between the model as a whole and the data are the R factor and the 'free R factor', R_{free}. Criteria for evaluating the local agreement of the model with electron density on a per-atom or per-residue basis are also available, and, more recently, access to more powerful computers has made it possible to compute the standard uncertainties of individual parameters, such as atomic coordinates or thermal factors.

Finally, the growing number of atomic resolution structures – primarily of proteins – is starting to provide a valuable source of much more precise information about the structures' geometrical and conformational properties. This should contribute to the improvement of standard values used in validation.

In this chapter, we present an overview of the different types of validation procedures applied to proteins and nucleic acids. We illustrate, in some detail, an approach to model validation based on atomic volumes embodied in the program *PROVE* and describe the package *SFCHECK*, which combines a set of criteria for evaluating the quality of the experimental data and the agreement of the model with the data.

21.2.2. Validating the geometric and stereochemical parameters of the model

21.2.2.1. Comparisons against standard values derived from crystals of small molecules

This concerns the validation of the covalent geometry of the atomic model. It involves comparing the bond distances and angles of the macromolecule against standard values and their associated uncertainties, derived from crystal structures of small organic molecules available in the Cambridge Structural Database, CSD (Allen *et al.*, 1979, 1983).

The standard values derived in this way are also used as restraints in crystallographic refinement programs, such as *XPLOR* (Brünger, 1992*a*) or the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994). As a result, the bond distances and angles of the final model usually agree well with their standard values, and the degree of scatter merely reflects the relative weight imposed on the various terms of the target function during refinement.

For proteins, the most commonly used standard values for the bond distances and angles are those compiled by Engh & Huber (1991) from molecular fragments in the CSD that most closely resemble chemical groups in amino acids. These parameters were shown to yield an improved description over that provided by the param19x.pro used in *XPLOR*, especially for the covalent geometry of aromatic rings in side-chain groups. It is noteworthy that these CSD-derived bond distances and angles can differ significantly from those used in molecular dynamics force fields, such as that of a recent version of *CHARMM* (MacKerell *et al.*, 1998). In these force fields, covalent-geometry parameters are obtained by a different strategy. They are optimized together with non-bonded parameters against a large body of available energy and structural data for a limited set of compounds representing amino-acid building blocks.

Protein-structure validation packages, such as *PROCHECK* (Laskowski *et al.*, 1993) and *WHAT IF* (Hooft, Vriend *et al.*, 1996), flag all bond distances and angles that deviate significantly from the database-derived reference values. This includes analysis of the deviations from planarity in aromatic rings and planar side-chain groups.

Similar checks are performed for the covalent geometry of atomic models of RNA or DNA oligo- and polynucleotides. Here, standard ranges for bond distances and angles are derived from crystal structures of nucleic acid bases, mononucleosides and mononucleotides in the CSD (Clowney *et al.*, 1996; Gelbin *et al.*, 1996). These values are used in validation procedures developed

Copyright © 2012 International Union of Crystallography

by the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) and in crystallographic refinement programs. For higher-resolution structures (better than 2.4 Å), a standard geometry, dependent on the sugar pucker conformation (C2′ *endo* or C3′ *endo*) (Parkinson *et al.*, 1996), is used.

Validation of the covalent geometry of the so-called 'hetero groups' (chemically modified monomer groups or small molecules that bind to macromolecules) is much more difficult. It therefore tends not to be routinely performed, and, as a result, the quality of the hetero groups in the models deposited in the Protein Data Bank (PDB) (Bernstein *et al.*, 1977; Berman *et al.*, 2000) varies widely.

The variety of the chemical structures of these molecules (the current release of the PDB contains about 2700 chemically distinct compounds) makes it difficult to archive them consistently, let alone to compile the dictionaries containing the required reference values in advance. Proper handling and verification of such groups require a comprehensive and rigorous description of the chemical components, as well as flexible means of deriving the appropriate reference geometries.

The development of systematic procedures for checking bond lengths and various torsion angles of hetero groups (Kleywegt & Jones, 1998) is a step in the right direction. Further progress should come, thanks in part to the recently adopted macromolecular Crystallographic Information File (mmCIF) format (Bourne *et al.*, 1997), which provides the necessary framework for a much more comprehensive and rigorous description of the molecular components. Using this description as the basis, automated tools for building 'customized' dictionaries of geometrical standards have been developed. One such tool is A LigAnd and Monomer Object Data Environment (*A LA MODE*) (Clowney *et al.*, 1999). It starts from a minimal topological description of a ligand or monomer component and performs the tasks required to construct the mmCIF component description. This includes querying the CSD, integration and book-keeping of database survey results, analysis and comparison of covalent geometry and stereochemistry, and the assembly of complex model structures from the results of multiple database surveys. Tools such as this considerably simplify the handling of small molecules at the refinement, validation and archiving stages.

21.2.2.2. *Comparisons against standard values derived from surveys of other macromolecules*

This involves computing a number of stereochemical, geometric and energy parameters from the atomic coordinates of the macromolecule and comparing them with standard ranges derived from high-quality crystal structures of other macromolecules. These standards represent the 'expected' properties, and the aim is to evaluate the quality of a model by measuring the extent to which it departs from these properties.

This evaluation is usually performed at the global level, in order to assess the quality of the structure as a whole, and on the local level, to identify specific regions with unusual properties. Such regions may represent genuine problems with the model or unusual conformations adopted for functional purposes, and it is sometimes difficult to distinguish between these two alternatives. The choice of reference structures from which the standards are derived is a crucial aspect of the approach, since both the mean and shape of the reference distributions may be affected by it.

21.2.2.2.1. *Validation of stereochemical and non-bonded parameters*

Morris *et al.* (1992) pioneered this type of validation for proteins. The software *PROCHECK* (Laskowski *et al.*, 1993), which implements and extends this approach, is described in detail in Chapter 21.4 of this volume. A very important evaluation criterion is the Ramachandran-plot quality, where the distribution of the backbone φ, ψ angles of a given protein structure is compared to that in high-quality structures. The comparison is performed both globally, by determining the proportion of the residues in favourable (core) regions of the plot, and locally, by the log-odds (G-factor) value, which measures how normal or unusual a residue's location is in the plot for a given residue type.

A similar strategy is used to evaluate other stereochemical parameters, such as the side-chain torsion angles (χ_1, χ_2, χ_3 *etc.*), the peptide bond torsion (ω), the C^α tetrahedral distortion, disulfide bond geometry and stereochemistry.

An evaluation of the backbone hydrogen-bonding energy is also performed, using the Kabsch & Sander (1983) algorithm, by comparison with distributions computed from high-resolution protein structures.

Other programs like *WHAT IF* (Hooft, Vriend *et al.*, 1996) perform similar evaluations. This program computes the expected φ, ψ distribution for each residue type from a data set of non-redundant high-quality structures and evaluates how the φ, ψ distribution of a given protein deviates from the expected values (Hooft *et al.*, 1997). A somewhat different version of this approach is proposed by Kleywegt & Jones (1996). *WHAT IF* also computes other quality indicators such as the number of buried unsatisfied hydrogen bonds or the extent of the overlap of van der Waals spheres ('clashes'). In addition, it verifies the orientation of His, Gln and Asn side chains, based on a hydrogen-bond network analysis, which also takes into account hydrogen bonds between symmetry-related molecules (Hooft, Sander & Vriend, 1996).

The very small fraction of structures (< 1.3%) for which only the C^α coordinates are deposited cannot be validated by the standard techniques. For these structures, two sets of parameters were shown to be useful (Kleywegt & Jones, 1996). They are the $C^\alpha - C^\alpha$ distances and a Ramachandran-like plot which displays for each residue the $C^\alpha_{i-1} - C^\alpha_i - C^\alpha_{i+1} - C^\alpha_{i+2}$ dihedral angle against the $C^\alpha_{i-1} - C^\alpha_i - C^\alpha_{i+1}$ angle. Deviations from the expected distributions of these parameters, computed from a set of high-quality complete protein structures, are used as quality indicators.

The validation of nucleic acid stereochemistry, in particular DNA, has a much shorter history. Only in recent years has the number of high-quality nucleic acid crystal structures become large enough to permit the derivation of reliable conformational trends. Schneider *et al.* (1997) derived ranges and mean values for the torsion angles of the sugar–phosphate backbone in helical DNA from a set of 96 oligodeoxynucleotide crystal structures. These ranges form the basis for the nucleic acid structure validation protocols currently implemented at the NDB.

21.2.2.2.2. *Validation using knowledge-based interaction potentials and profiles*

These methods represent a distinct set of approaches to the validation of the non-bonded and conformational parameters of the model. They involve computing the relative frequencies of residue–residue or atom–atom contacts from a set of high-quality protein structures and evaluating how the contacts in a given protein deviate from these standard frequencies. Most often, these frequencies are translated into potentials (energies) using the Boltzmann relation (Sippl, 1990), and these 'knowledge-based' potentials are used to score the structure (for a review, see Wodak & Rooman, 1993). The potentials that consider residue–residue interactions, as in the software *PROSA II* (Sippl, 1993), are usually quite crude since each residue is represented by a single interaction centre. They can therefore detect only gross errors in chain tracing or identify incorrectly modelled segments in an otherwise correct structure, but can not validate detailed atomic positions. The same limitation applies to procedures based on three-dimensional (3D) environment profiles (Eisenberg *et al.*, 1997). The latter consider the relative frequencies of finding each of the 20 amino acids in a given local 3D environment defined by the residue buried area, the ratio of polar *versus* non-polar neighbours and the secondary structure. The corresponding energies are used to score the compatibility of a structure with its amino-acid sequence in a manner similar to the residue–residue interaction potentials.

Finally, validation procedures based on the relative frequencies of atom–atom interactions in known protein structures have also been developed (Melo & Feytmans, 1997, 1998). These methods, consolidated in the software *ANOLEA*, are capable of identifying local errors and problems of sequence misalignment in protein structures built by homology modelling. In addition, energy Z scores computed with these potentials for whole protein structures correlate well with the resolution of the X-ray data, as shown below for the volume-based Z scores.

21.2.2.2.3. Deviations from standard atomic volumes as a quality measure for protein crystal structures

The observations that protein X-ray structures are at least as tightly packed as small-molecule crystals (Richards, 1974; Harpaz *et al.*, 1994) and that the packing density inside proteins displays very limited variation (Richards, 1974; Finney, 1975) suggest that atomic volumes or measures of atomic packing can be added to the list of parameters for assessing the quality of protein structures.

Packing and related measures have been used to compare structures of proteins derived by both X-ray diffraction and NMR spectroscopy. Ratnaparkhi *et al.* (1998) analysed pairs of protein structures for which both crystal and NMR structures were available. They found that the packing values of the NMR models displayed a much larger scatter than those of the corresponding crystal structures, suggesting that this is probably due to the fact that accurate values of the packing density cannot, at present, be obtained from NMR data. Similar conclusions were reached using measures of residue–residue contact area (Abagyan & Totrov, 1997).

Here, we describe the approach of Pontius *et al.* (1996), in which deviations from standard atomic volumes are used to assess the quality of a protein model, both overall and in specific regions.

The volumes occupied by atoms and residues inside proteins can be readily computed using the Voronoi method (Voronoi, 1908), first applied to proteins by Richards (1974) and Finney (1975). This method uses the atomic positions of the molecular model, and the volume assigned to each atom is defined as the smallest polyhedron created by the set of planes bisecting the lines joining the atom centre to those of its neighbours (Fig. 21.2.2.1).

The use of the classical Voronoi procedure is justified in the context of validation because it avoids the need to derive a consistent set of van der Waals radii for atoms in the system. Such sets are used by other volume-calculation methods in order to partition space more accurately (Richards, 1974, 1985; Gellatly & Finney, 1982). Assigning a consistent set of radii to protein atoms is, indeed, not straightforward due to the heterogeneity of the interactions within the protein (polar, ionic, non-polar) and the presence of a large variety of hetero groups.

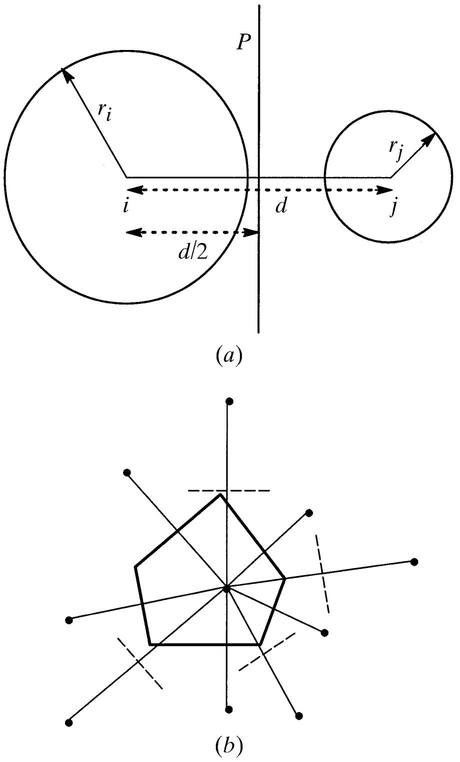


Figure 21.2.2.1
The Voronoi polyhedron. (*a*) Positioning of the dividing plane P between two atoms i and j, with van der Waals radii r_i and r_j, respectively, separated by a distance d. The plane P is positioned at $d/2$. (*b*) 2D representation of the Voronoi polyhedron of the central atom. This polyhedron is the smallest polyhedron delimited by all the dividing planes of the atom.

Structure-quality assessment based on volume calculations involves computing the atomic volumes in a subset of highly resolved and refined protein structures and analysing the distributions of these volumes for different atomic types, defined according to their chemical nature and bonded environment. These distributions define the expected ranges (mean and standard deviation) for the volume of each category of atoms. Atomic volumes in a given structure are then compared to the expected ranges, and statistically significant deviations from these ranges are flagged.

The program *PROVE* (Pontius *et al.*, 1996) implements such an approach using the analytic algorithms for volume and surface-area calculations encoded in *SurVol* (Alard, 1991). It computes for each atom i in a structure its volume Z score ($Z \text{ score} = \left|V_i^k - \overline{V^k}\right| / \sigma^k$), where the superscript k designates the particular atom type (*e.g.*, the C^α atom in a Leu residue), and $\overline{V^k}$ and σ^k are, respectively, the mean and standard deviation of the reference volume distribution for the corresponding atom type. These reference distributions are derived from a set of high-quality protein crystal structures using exactly the same calculation procedure (Pontius *et al.*, 1996).

Atoms with absolute Z scores > 3 are flagged as possible problem regions in the protein model, and residues containing such atoms are highlighted on graphical plots of the same type as those used by the *PROCHECK* program and on molecular

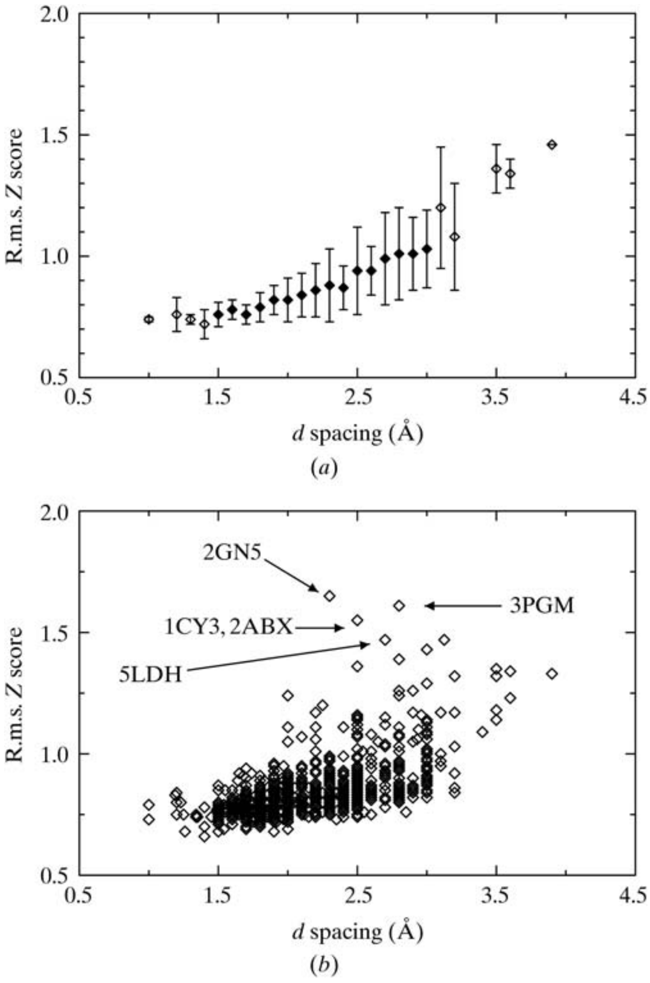


Figure 21.2.2.2
Atomic volume Z score r.m.s. variation with nominal resolution (d spacing) in 900 protein structures from the PDB. (*a*) Average of the r.m.s. volume Z score computed for structures having the same resolution (to within ±0.1 Å). The vertical bars indicate the magnitude of the standard deviations of the r.m.s. volume Z score in individual d-spacing bins. Graph points are derived from less than 10 structures (open diamonds) and from more than 10 structures (filled diamonds). (*b*) R.m.s. Z-score values as in (*a*), displayed for individual structures as a function of resolution. The five furthest outlier proteins are marked by their PDB codes.

models displayed using programs such as *Rasmol* (Sayle & Milner-White, 1995).

In addition to the validation of the local quality of the model, its overall quality can be assessed by the *root-mean-square volume Z score* of all its atoms (see Fig. 21.2.2.2 for definition). As for many stereochemical global quality indicators, this Z score shows good correlation with the nominal resolution (d spacing) of the crystallographic data, as illustrated in Fig. 21.2.2.2(*a*). This figure also shows that Z-score ranges can be defined for each resolution interval. The Z scores of individual proteins that lie outside these intervals may be indicative of 'problem' structures. This is clearly the case for the two proteins 2ABX and 2GN5, whose Z scores are much higher than average (Fig. 21.2.2.2*b*).

Since the Voronoi volume of solvent-accessible atoms cannot be defined, because these atoms are not completely surrounded by other atoms, only completely buried atoms are scored.

The current version of *PROVE* is unable to measure the deviations from standard volumes for atoms in nucleic acids or hetero groups, simply because of the lack of reference volumes for these structures. This should change in the near future, at least for nucleic acids, thanks to the growing number of high-quality nucleic acid crystal structures from which standard volume ranges could be readily derived.

21.2.3. Validation of a model *versus* experimental data

By far the most important measure of the quality of a given atomic model is its agreement with the experimental data. This type of validation is geared towards detecting systematic errors, which determine the overall accuracy of the model, and random errors, which affect the precision of the model.

Systematic errors are difficult to detect even in highly refined structures, especially at lower resolution. The most commonly used measures of the agreement between the atomic coordinates and the X-ray data are the classical R factor and the 'free R factor' ($R_{\rm free}$) (Brünger, 1992*b*). The latter is based on standard statistical cross-validation techniques (Brünger, 1997) and is therefore less amenable to manipulation, such as leaving out weak data or over-fitting the data with too many parameters. Currently, nearly half of the publications on macromolecular structures report $R_{\rm free}$ values, an indication that its use is becoming more widespread. So far, however, there are no clear guidelines indicating what an 'acceptable' $R_{\rm free}$ value should be (Kleywegt & Brünger, 1996).

An expression for estimating the expected $R_{\rm free}$ value has been proposed (see Dodson *et al.*, 1996) and used to assess the significance of the drop in $R_{\rm free}$ during refinement. Accurate expressions for the expected ratio of $R_{\rm free}$ to R (the $R_{\rm free}$ ratio) have also been derived theoretically (Tickle *et al.*, 1998). This ratio seems to be independent of random errors and can be used to detect systematic errors at the convergence of the least-squares refinement. The remaining problem is to determine what the precision of $R_{\rm free}$ or the $R_{\rm free}$ ratio should be. In other words, if the $R_{\rm free}$ ratio differs from the expected value, when is the difference significant? This requires knowing the variance of these parameters. Estimating the precision of $R_{\rm free}$ can be done empirically by performing repeated refinements of the same structure with different sets of reflections removed (Brünger, 1997). From such analysis, a useful approximation to the $R_{\rm free}$ precision was suggested to be the ratio $R_{\rm free}/(n)^{1/2}$, where n is the number of reflections in the test set.

Evaluating the precision of the refined parameters, that is, the atomic coordinates and the temperature or B factors, is a different matter. In small-molecule crystallography, the standard uncertainty (s.u.) of the parameters can be computed from the variance–covariance matrix, obtained by inverting the full normal-equations matrix (Cruickshank, 1965). This can, in principle, also be done for the parameters of macromolecules. However, the number of second derivatives to be computed and the size of the matrix to be inverted are so large that this task is too time consuming to be performed routinely. This is gradually changing, however. An increasing number of proteins structures, primarily those solved at atomic resolution, have their s.u.'s computed in this manner (Deacon *et al.*, 1997; Harata *et al.*, 1998). A program often used for this purpose is *SHELXL* (Sheldrick & Schneider, 1997), a well known refinement software package for small molecules that has recently been extended to proteins. Availability of s.u.'s can determine the dependence of the precision of the atomic coordinates on various factors, such as the resolution, the atomic number, and the number and types of restraints used during refinement (Tickle *et al.*, 1998).

Other methods for determining the relative precision of atoms in macromolecular structures involve calculating the agreement between the model and the electron density in specific regions.

Table 21.2.3.1
Parameters computed for the analysis of the structure-factor data

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
Completeness (%)	Percentage of the expected number of reflections for the given crystal space group and resolution	
B_overall (Patterson)	$8\pi^2\sigma_{\rm Patt}/(2)^{1/2}$†	Overall B factor
R_stand(F)	$\langle\sigma(F)\rangle/\langle F\rangle$‡	Uncertainty of the structure-factor amplitudes
Optical resolution	$(\sigma^2_{\rm Patt}+\sigma^2_{\rm sph})^{1/2}$† §	Expected minimum distance between two resolved atomic peaks
Expected optical resolution	Optical resolution computed considering all reflections	
CC_F	$\dfrac{\langle F_{\rm obs}F_{\rm calc}\rangle-\langle F_{\rm obs}\rangle\langle F_{\rm calc}\rangle}{\left[(\langle F^2_{\rm obs}\rangle-\langle F_{\rm obs}\rangle^2)(\langle F^2_{\rm calc}\rangle-\langle F_{\rm calc}\rangle^2)\right]^{1/2}}$	Correlation coefficient between the observed and calculated structure-factor amplitudes
S	$\left\{\dfrac{\sum(F_{\rm obs}f_{\rm cutoff})^2}{\sum\left[F_{\rm calc}\exp(-B^{\rm overall}_{\rm diff}s^2)f_{\rm cutoff}\right]^2}\right\}^{1/2}$ ¶	Factor applied to scale $F_{\rm calc}$ to $F_{\rm obs}$
$f_{\rm cutoff}$	$1-\exp(-B_{\rm off}s^2)$††	Function applied to obtain a smooth cutoff for low-resolution data

† $\sigma_{\rm Patt}$ is the standard deviation of the Gaussian fitted to the Patterson origin peak. ‡ F is the structure-factor amplitude, and $\sigma(F)$ is the structure-factor standard deviation. The brackets denote averages. § $\sigma_{\rm sph}$ is the standard deviation of the spherical interference function, which is the Fourier transform of a sphere of radius $1/d_{\rm min}$, with $d_{\rm min}$ being the minimum d spacing. ¶ $B^{\rm overall}_{\rm diff}=B^{\rm overall}_{\rm obs}-B^{\rm overall}_{\rm calc}$ is added to the calculated overall B factor, $B_{\rm overall}$, so as to make the width of the calculated Patterson origin peak equal to the observed one; s is the magnitude of reciprocal-lattice vector. †† $B_{\rm off}=4d^2_{\rm max}$, where s and $d_{\rm max}$, respectively, are the magnitude of the reciprocal-lattice vector and the maximum d spacing.

The newer approach by Zhou *et al.* (1998) is related to the real-space R factor of Jones *et al.* (1991), but differs from it by the way in which the electron density is computed (Chapman, 1995).

As our understanding of the factors that govern the systematic errors in macromolecular crystallography increases and our ability to detect random errors improves, the possibility of devising systematic and possibly more automatic protocols for assessing the agreement between the model and the data will emerge.

In what follows, we describe the software package *SFCHECK* (Vaguine *et al.*, 1999), which can be regarded as a first attempt in this direction. This software computes and summarizes many of the commonly used measures for evaluating the quality of the structure-factor data and the agreement of the model with these data.

We summarize the tasks performed and the quality indicators computed by *SFCHECK* and briefly illustrate how this software can be used to evaluate individual structures and survey different structures.

21.2.3.1. A systematic approach using the SFCHECK software

21.2.3.1.1. Tasks performed by SFCHECK

21.2.3.1.1.1. Treatment of structure-factor data and scaling

SFCHECK reads in the structure-factor data written in mmCIF format. It then performs the following operations: Reflections are excluded if they are systematically absent, negative, or have flagged σ values (99.9). Equivalent reflections are merged. The amplitudes of missing reflections are approximated by taking the average value for the corresponding resolution shell.

From the model coordinates read from the PDB (or mmCIF) atomic coordinates file, *SFCHECK* calculates structure factors and scales them to the observed structure factors. The scaling factor, S, is computed using a smooth cutoff for low-resolution data (Vaguine *et al.*, 1999) (Table 21.2.3.1). This involves the calculation of the observed and calculated overall B factors from the standard deviations of the Gaussian fitted to the Patterson origin peaks [see Table 21.2.3.1 and Vaguine *et al.* (1999)]. In addition, *SFCHECK* also estimates the overall anisotropy of the data, following the approach of Sheriff & Hendrickson (1987), and applies the anisotropic scaling after the Patterson scaling is performed (Murshudov *et al.*, 1998).

To assess the quality of the structure-factor data, the program computes four additional quantities (see Table 21.2.3.1 for details): the completeness of the data, the uncertainty of the structure-factor amplitudes, the optical resolution and the expected optical resolution. The latter two quantities represent the expected minimum distance between two resolved atomic peaks in the electron-density map when the latter is computed with the set of reflections specified by the authors and with all the reflections, respectively.

21.2.3.1.1.2. Global agreement between the model and experimental data

To evaluate the global agreement between the atomic model and the experimental data, the program computes three classical quality indicators: the R factor, $R_{\rm free}$ (Brünger, 1992*b*) and the correlation coefficient CC_F between the calculated and observed structure-factor amplitudes (Table 21.2.3.1). The R factor is computed using all the reflections considered (except those approximated by their average value in the corresponding resolution shell) and applying the same resolution and σ cutoff as those reported by the authors. $R_{\rm free}$ is computed using the subset of reflections specified by the authors. In addition, the R factor is evaluated using the 'non-free' subset of reflections (those not used to compute $R_{\rm free}$). The correlation coefficient is computed using all reflections from the reported high-resolution limit, applying the smooth low-resolution cutoff (see Table 21.2.3.1) but no σ cutoff.

Table 21.2.3.2
Estimation of errors in atomic coordinates

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
$\sigma(x)$	$\dfrac{\sigma(\text{slope})}{\text{curvature}}$†	Standard deviation of the atomic coordinates following Cruickshank (1949) for the minimal and maximal errors (Vaguine *et al.*, 1999)
σ(slope) for maximal error	$\dfrac{2\pi\{\sum[h^2(F_{\text{obs}}-F_{\text{calc}})^2]\}^{1/2}}{V_{\text{unit cell}}a}$‡	Expression for σ(slope) in the expected maximal error following Cruickshank (1949)
Curvature	$\dfrac{2\pi\sum(h^2F_{\text{obs}})}{V_{\text{unit cell}}a^2}$	Expression for the curvature following Murshudov *et al.* (1997)
σ(slope) for minimal error	$\dfrac{2\pi^2\{\sum[h^2\sigma(F_{\text{obs}})^2]\}^{1/2}}{V_{\text{unit cell}}a}$§	Expression for σ(slope) in the expected minimal error, following Cruickshank (1949)
DPI	$\sigma(x)=\left(\dfrac{N_{\text{atoms}}}{N_{\text{obs}}-4N_{\text{atoms}}}\right)^{1/2}c^{-1/3}d_{\min}R$¶	Atomic coordinate error estimate following Cruickshank (1996)

† σ(slope) and curvature are the slope and curvature of the electron-density map at the atomic centre, in the x direction, for spherically symmetric peaks; $\sigma(x)\simeq\sigma(y)\simeq\sigma(z)$. ‡ a is the crystal unit-cell length, h is the Miller index and $V_{\text{unit cell}}$ the unit-cell volume. § $\sigma(F_{\text{obs}})$ is the standard deviation of the structure-factor amplitude. ¶ c is the structure-factor data completeness expressed as a fraction (0–1), R is the conventional R factor, N_{atoms} is the total number of atoms in the unit cell, N_{obs} is the total number of observed reflections and $d_{\min}$ is the minimum d spacing.

21.2.3.1.1.3. Estimations of errors in atomic positions

The errors associated with the atomic positions are expressed as standard deviations (σ) of these positions. *SFCHECK* computes three different error measures. One is the original error measure of Cruickshank (1949). The second is a modified version of this error measure, in which the difference between the observed and calculated structure factors is replaced by the error in the experimental structure factors. The first two error measures are the expected maximal and minimal errors, respectively, and the third measure is the diffraction-component precision indicator (DPI). The mathematical expressions for these error measures are given in Table 21.2.3.2, and further details can be found in Vaguine *et al.* (1999).

21.2.3.1.1.4. Local agreement between the model and the experimental data

In addition to the global structure quality measures, *SFCHECK* also determines the quality of the model in specific regions. Several quality estimators can be calculated for each residue in the macromolecule and, whenever appropriate, for solvent molecules and groups of atoms in ligand molecules. These estimators are the normalized atomic displacement (Shift), the correlation coefficient between the calculated and observed electron densities (Density correlation), the local electron-density level (Density index), the average B factor (B-factor) and the connectivity index (Connect), which measures the local electron-density level along the molecular backbone. These quantities are computed for individual atoms and averaged over those composing each residue or group of atoms [see Table 21.2.3.3 and Vaguine *et al.* (1999) for details].

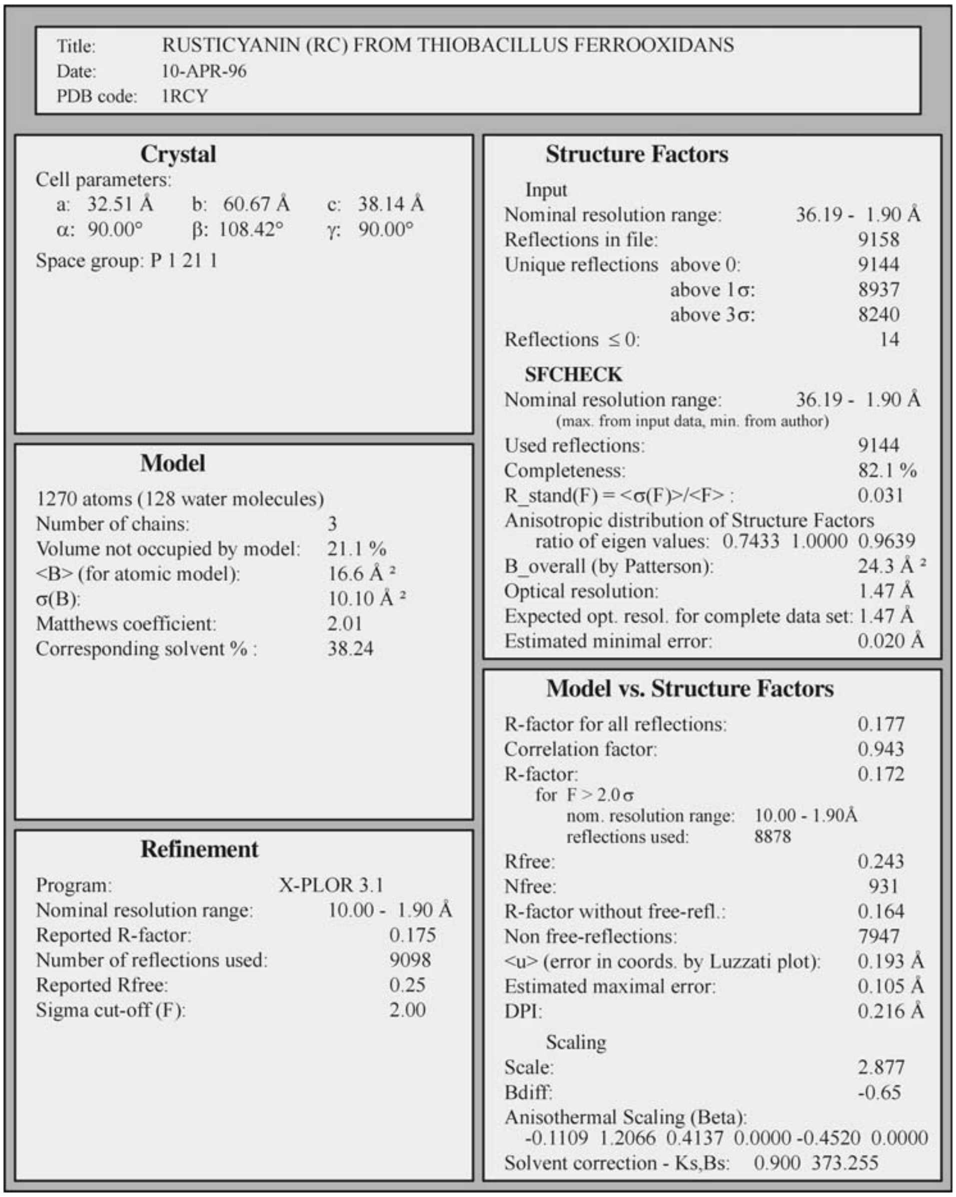
21.2.3.1.2. Evaluation of individual structures

Figs. 21.2.3.1–21.2.3.3 summarize the analysis carried out by *SFCHECK* on the protein rusticyanin from *Thiobacillus ferrooxidans* (1RCY) (Walter *et al.*, 1996). Fig. 21.2.3.1 displays the numerical results from the analysis of the structure-factor data and from the evaluation of the global agreement between the model and the data. The R-factor and R_{free} values, computed by *SFCHECK* (Model vs. Structure Factors panel) using the identical reflection subset to that reported by the authors (Refinement panel), show negligible differences with the reported values. These differences are 0.175 *versus* 0.172 for the R factor and 0.25 *versus* 0.243 for R_{free}. The small R-factor difference may stem from the fact that *SFCHECK* considers a somewhat different number of reflections (9144) than the authors (9098), although it uses the same d-spacing range and σ cutoff as those reported.

The information in Figs. 21.2.3.1 and 21.2.3.2 allows one to make some judgement about the quality of the structure-factor data for this protein. The relatively high resolution of this structure (1.9 Å) is accompanied by limited data completeness (82.1%). The Rstand(F) plot on the same graph shows, furthermore, a decrease in quality of the high-resolution data (2.2–1.9 Å). The average radial completeness plot (bottom left-hand plot of Fig. 21.2.3.2) allows one to identify the regions in reciprocal space with incomplete data.

Fig. 21.2.3.3 presents the *SFCHECK* analysis of the local agreement of the model with the electron density for 1RCY. The shift plot shows that both backbone and side-chain shifts are of comparable size, with several residues (1, 2, 16, 25) displaying shifts as high as 0.16 Å. The density correlation is excellent throughout the entire molecule, except for residues 2, 16 and 29, which display poorer correlation. In particular, the side chains of these residues seem to be more poorly defined in the electron-density map. The backbone density index plunges in a few regions, notably at the N-terminus (residues 5–7) and in the segments comprising residues 25–30 and 68–70. The side chains display, in general, a poorer density index than the backbone, with some regions (for example, residues 5–7, 23–30, 58–60) displaying rather low density indices. The same segments also display higher backbone and side-chain B factors. The backbone Connect parameter is, on the other hand, quite good throughout, except for residues 5–7 and 28–29 (Fig. 21.2.3.3).

Water molecules (labeled w in the *SFCHECK* output) are also evaluated. The relevant plots for these molecules are those of the Shift, Density index and B factor parameters. The first 50 or so water molecules in the list (appearing sequentially along the plot from left to right) tend to display a higher density index and lower B factors (< 30 Å^2) than the following molecules in the list.



```
Title:      RUSTICYANIN (RC) FROM THIOBACILLUS FERROOXIDANS
Date:       10-APR-96
PDB code:   1RCY

Crystal
Cell parameters:
  a:  32.51 Å      b:  60.67 Å     c:  38.14 Å
  α:  90.00°       β:  108.42°     γ:  90.00°
Space group: P 1 21 1

Model
1270 atoms (128 water molecules)
Number of chains:                3
Volume not occupied by model:    21.1 %
<B> (for atomic model):          16.6 Å ²
σ(B):                            10.10 Å ²
Matthews coefficient:            2.01
Corresponding solvent % :        38.24

Refinement
Program:                         X-PLOR 3.1
Nominal resolution range:           10.00 -  1.90 Å
Reported R-factor:                           0.175
Number of reflections used:                  9098
Reported Rfree:                              0.25
Sigma cut-off (F):                           2.00

Structure Factors
  Input
Nominal resolution range:            36.19 -  1.90 Å
Reflections in file:                          9158
Unique reflections  above 0:                  9144
                    above 1σ:                 8937
                    above 3σ:                 8240
Reflections  ≤ 0:                               14
  SFCHECK
Nominal resolution range:            36.19 -  1.90 Å
      (max. from input data, min. from author)
Used reflections:                             9144
Completeness:                                 82.1 %
R_stand(F) = <σ(F)>/<F> :                     0.031
Anisotropic distribution of Structure Factors
    ratio of eigen values:  0.7433  1.0000  0.9639
B_overall (by Patterson):                     24.3 Å ²
Optical resolution:                           1.47 Å
Expected opt. resol. for complete data set: 1.47 Å
Estimated minimal error:                      0.020 Å

Model vs. Structure Factors
R-factor for all reflections:                 0.177
Correlation factor:                           0.943
R-factor:                                     0.172
    for  F > 2.0σ
         nom. resolution range:    10.00 - 1.90Å
         reflections used:         8878
Rfree:                                        0.243
Nfree:                                         931
R-factor without free-refl.:                  0.164
Non free-reflections:                         7947
<u> (error in coords. by Luzzati plot):       0.193 Å
Estimated maximal error:                      0.105 Å
DPI:                                          0.216 Å
     Scaling
Scale:                                        2.877
Bdiff:                                        -0.65
Anisothermal Scaling (Beta):
  -0.1109  1.2066  0.4137  0.0000 -0.4520  0.0000
Solvent correction - Ks,Bs:    0.900  373.255
```

Figure 21.2.3.1

Typical *SFCHECK* output in PostScript format, illustrated for the protein rusticyanin from *Thiobacillus ferrooxidans* (1RCY) (Walter *et al.*, 1996). Summary panels displaying the numerical results from the analysis of the deposited structure-factor data and from the evaluation of the global agreement between the model and these data. The top elongated panel lists the PDB title record, deposition date and PDB code. The Crystal panel summarizes the crystal parameters, provided by the authors, as read from the model input files. The Model and Refinement panels list the information provided by the authors on the model and the refinement procedure, respectively. This information is read from the PDB coordinates entry. The Structure Factors panel summarizes the information on the deposited structure-factor data (Input section) and on the data used and criteria computed by *SFCHECK* (SFCHECK section). The numbers given under 'Anisotropic distribution of Structure Factors' are the ratios of the eigenvalues of the symmetric anisotropic thermal tensor to the maximum eigenvalues. The Model vs. Structure Factors panel summarizes the results of the verifications made by *SFCHECK*. The values listed under 'Anisothermal Scaling (Beta)' are those of the overall anisotropic thermal tensor ($b_{11}, b_{12}, b_{13}, b_{22}, b_{23}, b_{33}$). The meanings of other listed quantities are either self-explanatory or are described in the text.

They thus seem to be more reliably positioned than subsequent molecules, whose density indices sometimes drop perilously. A steady climb of the B factors is also apparent as one goes down the list of water molecules. The analysis of the density indices and B factors of individual water molecules performed by *SFCHECK* could be a very useful guide in investigations of the properties of crystallographic water molecules and their interactions with protein atoms.

21.2.3.1.3. Quality assessment based on surveys across structures

21.2.3.1.3.1. Assessing the quality of a structure as a whole

As for the evaluation of the geometric and stereochemical parameters of the model, surveying the same quality indicators across many structures is crucial. It allows one to establish the ranges of expected values for each indicator and to identify structures with unexpected features – those for which the values of one or more quality indicators are outside their standard range.

The global quality indicators computed by *SFCHECK* are the nominal resolution (d spacing), the R factor, $R_{\rm free}$, the minimal and maximal errors in atomic positions, the DPI, and the correlation coefficient CC_F. Another type of global quality indicator can be obtained by computing the average values of local quality measures across a given structure. This can be done for the per-residue (or per-group) atomic displacement and the Density correlation and B factor parameters as well as for the Density index and Connect parameters.

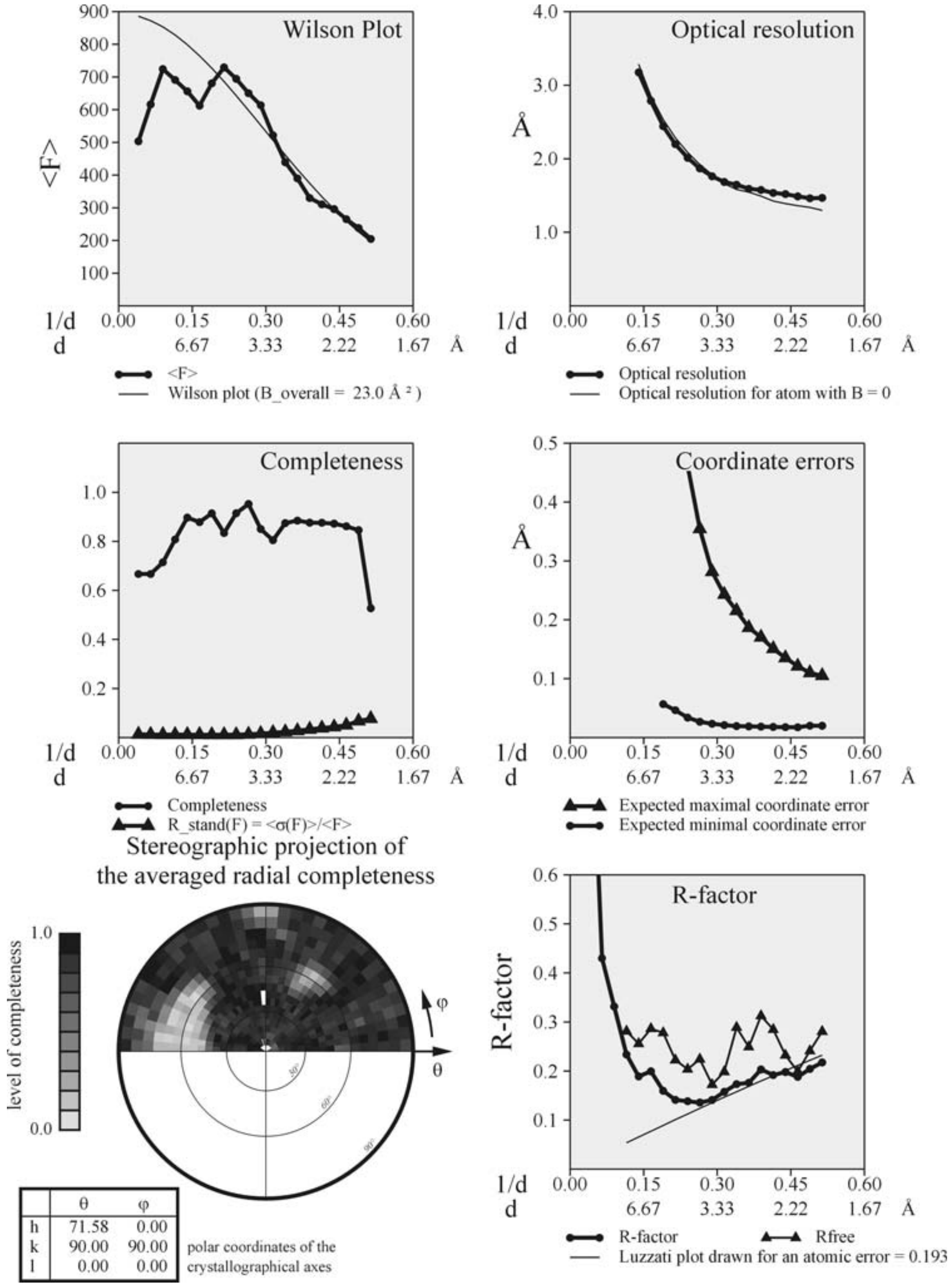


Figure 21.2.3.2

Graphical output from the *SFCHECK* analysis of global characteristics of the structure-factor data and the model agreement with those data for the same structure as in Fig. 21.2.3.1. From left to right and top to bottom: the Wilson plot; the behaviour of the optical resolution as a function of the nominal resolution (*d* spacing); the data completeness and structure-factor standard error as a function of the *d* spacing; the maximal and minimal coordinate error dependence on *d* spacing; a stereographic projection of the averaged radial structure-factor data completeness; and, finally, the *R*-factor dependence and Luzzati plots for a given atomic error.

Many of the geometric and stererochemical quality indicators vary as a function of resolution – some linearly and some not (Laskowski *et al.*, 1993). This is also the case for most of the global quality indicators described here. Examples of this dependence are given in Fig. 21.2.3.4, which shows how the correlation coefficient, the maximal error, the average atomic displacement and average density index vary as a function of resolution in the 104 nucleic acid structures surveyed. This variation is approximately linear for all four parameters. The density correlation and average density index decrease, whereas the maximal error and average atomic displacements increase, as the resolution gets poorer. In all four plots of Fig. 21.2.3.4, the points tend to display significant scatter as the *d* spacing increases, and at least three points, corresponding to the same three structures, appear as outliers in all plots. These structures also appear as outliers in the analysis of other parameters. A closer examination revealed that in the vast majority of the cases, the abnormal behaviour of these structures could be traced back to problems with data formats or errors that occurred during data deposition and entry processing.

As the number of structures with deposited structure-factor data becomes large enough, plots such as those of Fig. 21.2.3.4 could be used to define the expected range of values for a quality indicator in a structure determined at a given resolution or refined under given conditions. Structures yielding quality indicators outside this range could then be identified as unusual on a more solid statistical basis.

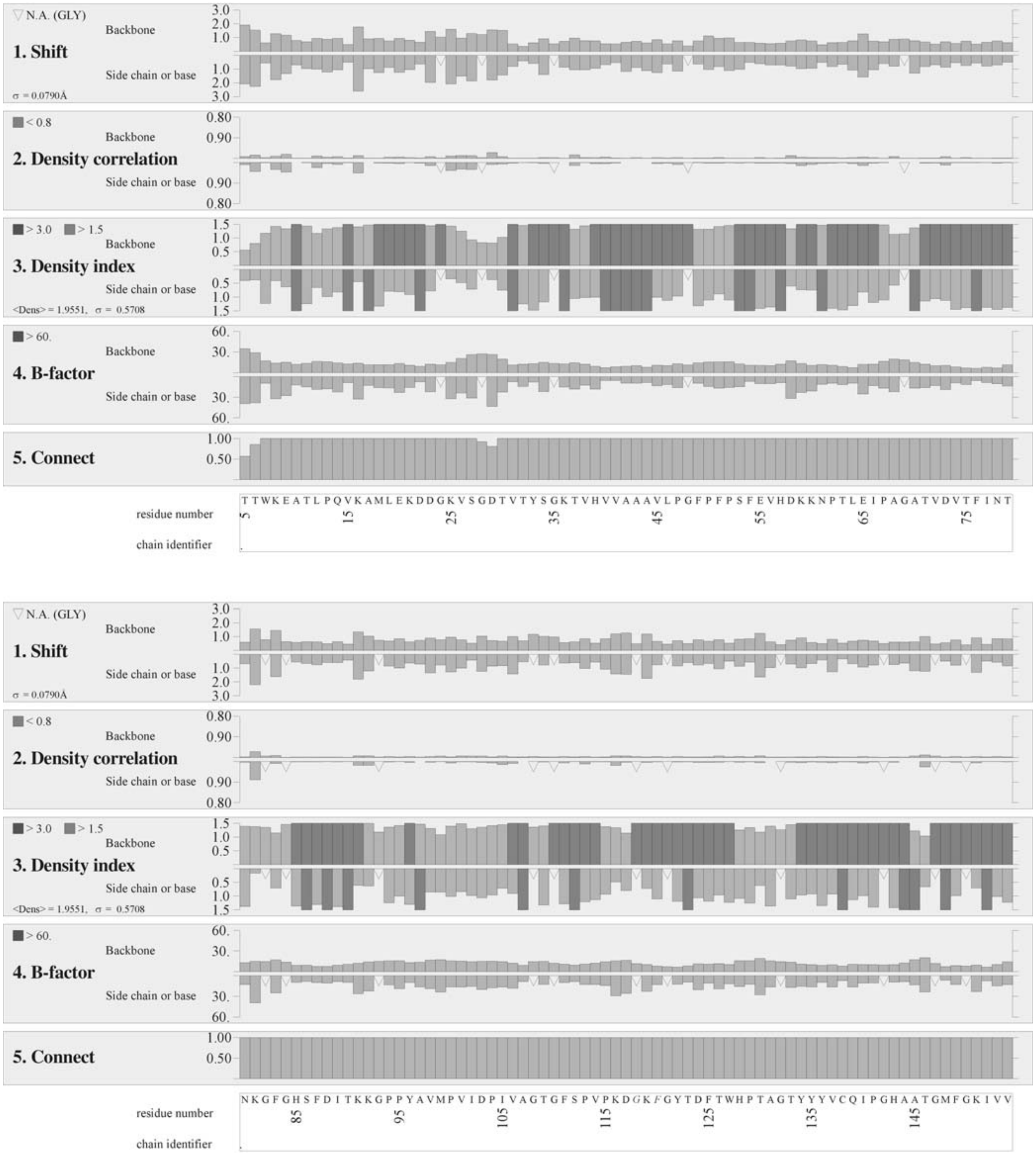


Figure 21.2.3.3

SFCHECK evaluation summary of the local agreement between the model and the electron density for the same structure as in Fig. 21.2.3.1. Five criteria are plotted for each residue of the macromolecule (designated by its one-letter code), as well as for each solvent molecule (w), or hetero group. These criteria are: (1) Shift, (2) Density correlation, (3) Density index, (4) B factor, (5) Connect. The definitions of these criteria are given in the text. Note that the values of the Connect parameter are truncated to a maximum of 1. The *SFCHECK* output shown in Figs. 21.2.3.1–21.2.3.3 was generated using routines from *PROCHECK* kindly provided by R. Laskowski.

21.2.3.1.3.2. Assessing the quality in specific regions of a model

The main purpose for computing the four local quality measures, the B factor, the Density index, the atomic displacement (Shift) and the Density correlation (Table 21.2.3.3), is to identify problem regions in a model. In order to do this effectively, it is necessary to evaluate the degree of redundancy between these measures and to establish the standard ranges for their values. The latter task, in particular, is not straightforward since it depends crucially on the quality of the experimental data and biases introduced by the scaling procedure and refinement protocol. In this regard, several issues are presently still under investigation.

A preliminary investigation of the mutual relations between the above-mentioned local measures has been performed in several protein and nucleic acid structures taken individually.

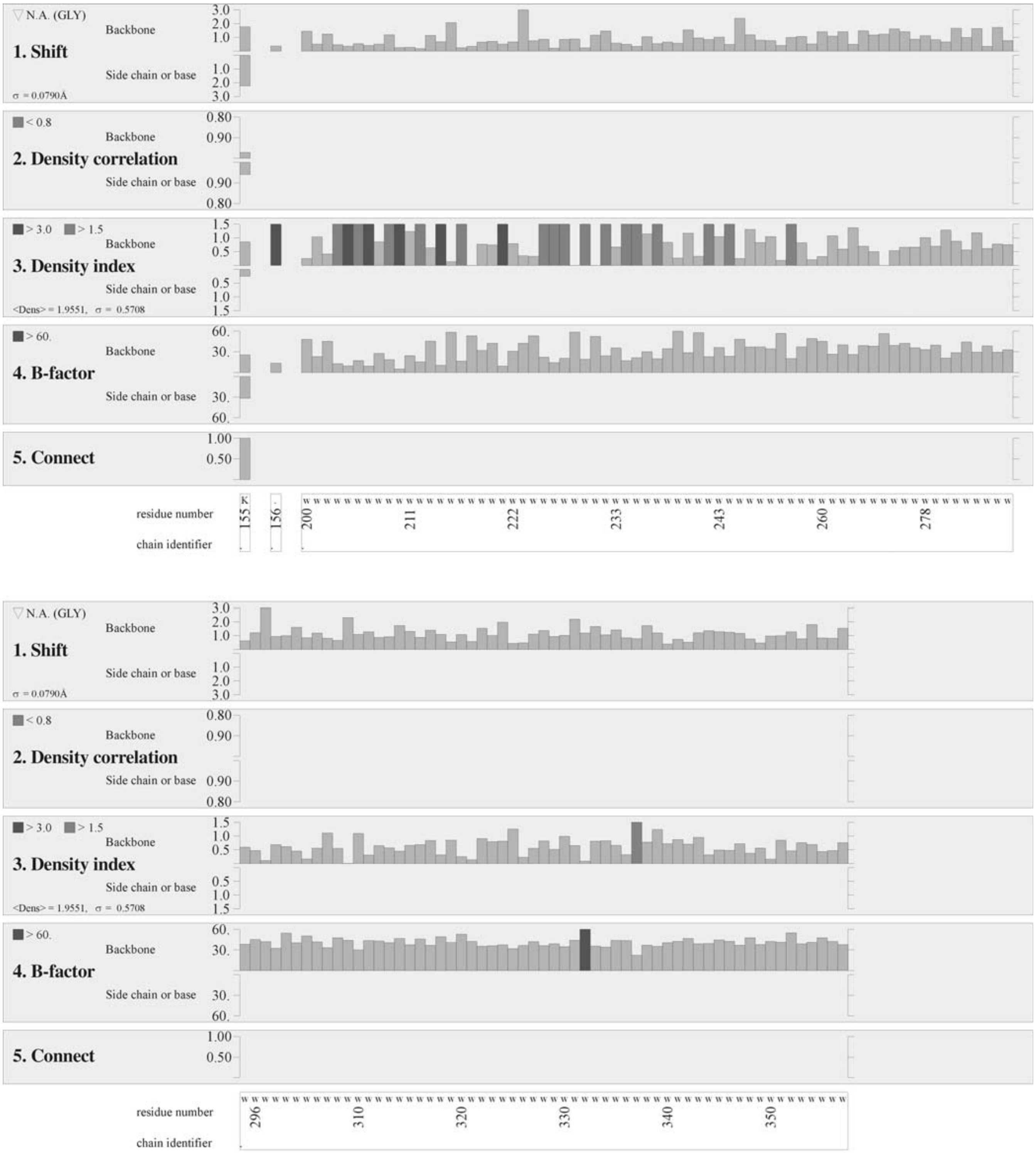


Figure 21.2.3.3 (continued)

This shows that the B factor is strongly correlated with the density index, as illustrated in Fig. 21.2.3.5(a), and to a lesser extent with the atomic displacement (Fig. 21.2.3.5b). A weaker correlation was detected between the latter three measures and the residue density correlation (data not shown).

Analyses across structures could, in principle, be carried out for all four local measures computed by *SFCHECK*, provided these measures are not subject to systematic biases due to differences in scaling procedures and refinement practices. Such biases are, however, well known for the B factors of individual atoms or residues. This is illustrated in Fig. 21.2.3.6(a). This figure plots, side-by-side, the average residue B factors in 21 protein structures determined at different d spacings. It shows that for proteins determined at poorer resolution (d spacing above 2 Å), the B factors of different structures are systematically shifted relative to one another. Such systematic shifts are much smaller for structures determined at 2 Å resolution or better (Fig. 21.2.3.6a). This is not surprising, since in lower-resolution structures, $N_{\text{refl}}/N_{\text{atoms}}$ is often too low (< 4) to yield meaningful values for the B factors.

Interestingly, the residue Density index, a very different parameter from the B factor, which measures the level of electron density at the atomic positions, does not display the systematic shifts observed for the B factors (Fig. 21.2.3.6b), despite the fact

Table 21.2.3.3
Parameters computed by *SFCHECK* to assess the quality of the model in specific regions

The first column lists the parameter, the second column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters when warranted.

Parameter	Formula/definition	Meaning
Shift	$(1/N\sigma)\sum_{i}^{N}\Delta_i$, with $\Delta_i = (\text{gradient}_i/\text{curvature}_i)$†	Normalized average atomic displacement computed over a group of atoms or residue; reflects the tendency of the group of atoms to move from their current position
Density correlation	$\dfrac{\sum\rho_{\rm calc}(x_i)[2\rho_{\rm obs}(x_i)-\rho_{\rm calc}(x_i)]}{\left(\left[\sum\rho^2_{\rm calc}(x_i)\right]\left\{\sum[2\rho_{\rm obs}(x_i)-\rho_{\rm calc}(x_i)]^2\right\}\right)^{1/2}}$‡	Electron density correlation coefficient computed over a group of atoms or residue; reflects the local agreement of the model with the electron density
Density index	$\left[\prod\rho(x_i)\right]^{1/N}/\langle\rho\rangle_{\text{all atoms}}$§	Reflects the level of the electron density for a group of atoms; is a local measure of the density level
Connect		Same as Density index, but considering only backbone atoms¶

† Gradient$_i$ is the gradient of the $F_{\rm obs}-F_{\rm calc}$ map with respect to the atomic coordinates, curvature$_i$ is the curvature of the model map computed at the atomic centre (see Agarwal, 1978), N is the number of atoms in the group considered and σ is the standard deviation of the Δ_i values computed in the structure. ‡ $\rho_{\rm calc}(x_i)$ and $\rho_{\rm obs}(x_i)$ are, respectively, the electron density computed from calculated and observed structure-factor amplitudes at the atomic centre. The summation is performed over all the atoms in the group considered. For polymer residues, D_corr is computed separately for backbone and side-chain atoms. For the calculation of the electron density at the atomic centre, see Vaguine *et al.* (1999). § $[\prod\rho(x_i)]^{1/N}$ is the geometric mean of the $2F_{\rm obs}-F_{\rm calc}$ electron density of the atom subset considered and $\langle\rho\rangle_{\text{all atoms}}$ is the average electron density of the atoms in the structure. For water molecules or ions which are represented by a unique atom, the above expression reduces to the ratio $\rho(x_i)/\langle\rho\rangle_{\text{all atoms}}$. ¶ Backbone atoms are N, C, C$^{\alpha}$ for proteins and P, O5′, C5′, C3′, O3′ for nucleic acids.

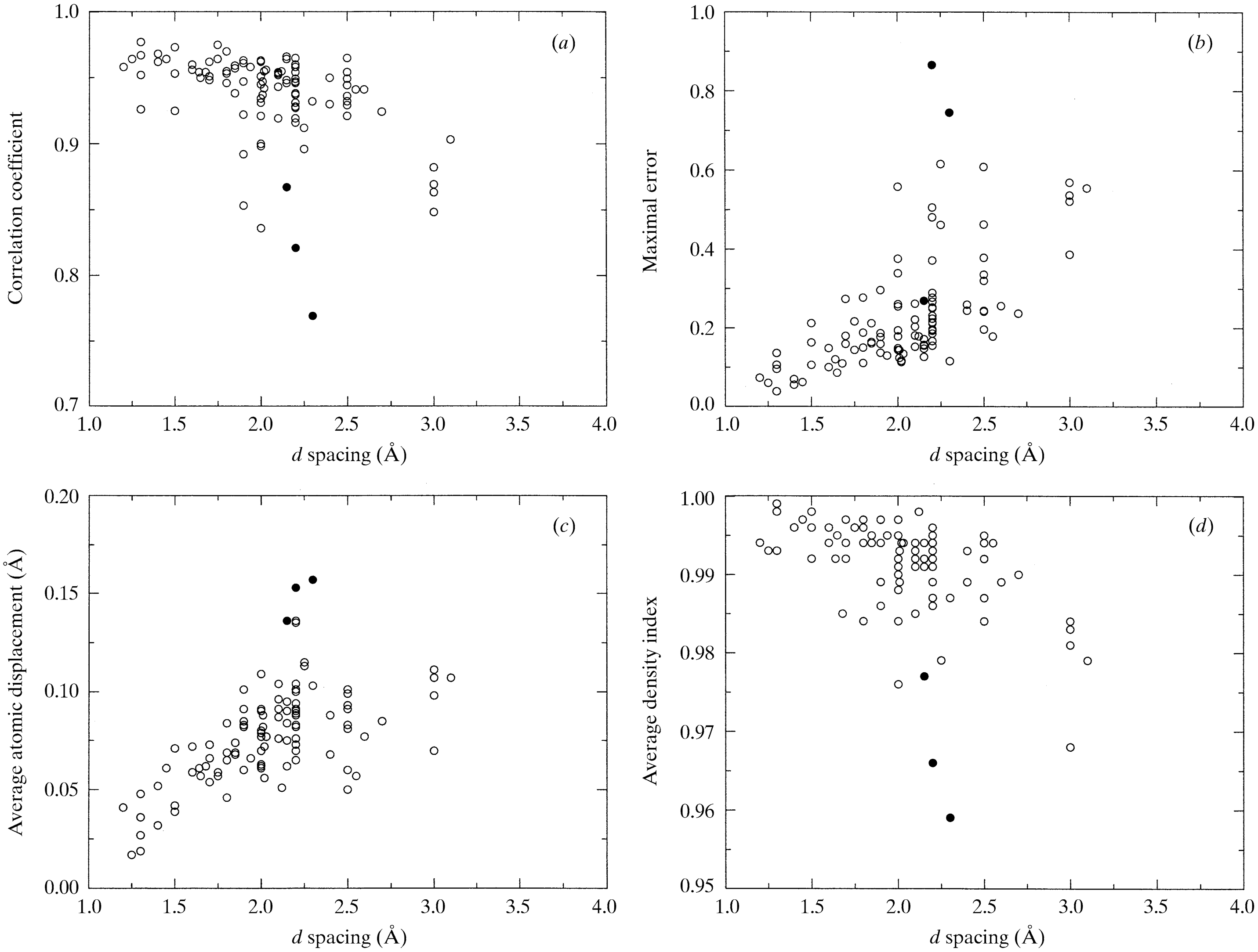


Figure 21.2.3.4
Variation of global quality indicators with the nominal resolution (d spacing) of the crystallographic data. The following quality indicators were computed by *SFCHECK* for each of the 104 nucleic acid crystal structures considered in the study of Das *et al.* (2001): (*a*) correlation coefficient, (*b*) maximal error, (*c*) average atomic displacement and (*d*) average density index. For the meaning of the various quantities see Table 21.2.3.2. The three structures for which the reported and re-computed R factors differ by more than 10% are highlighted as black circles. The NDB (PDB) codes for these structures are ADFB72 (256D), ADF073 (257D) and ADJ081 (320D).

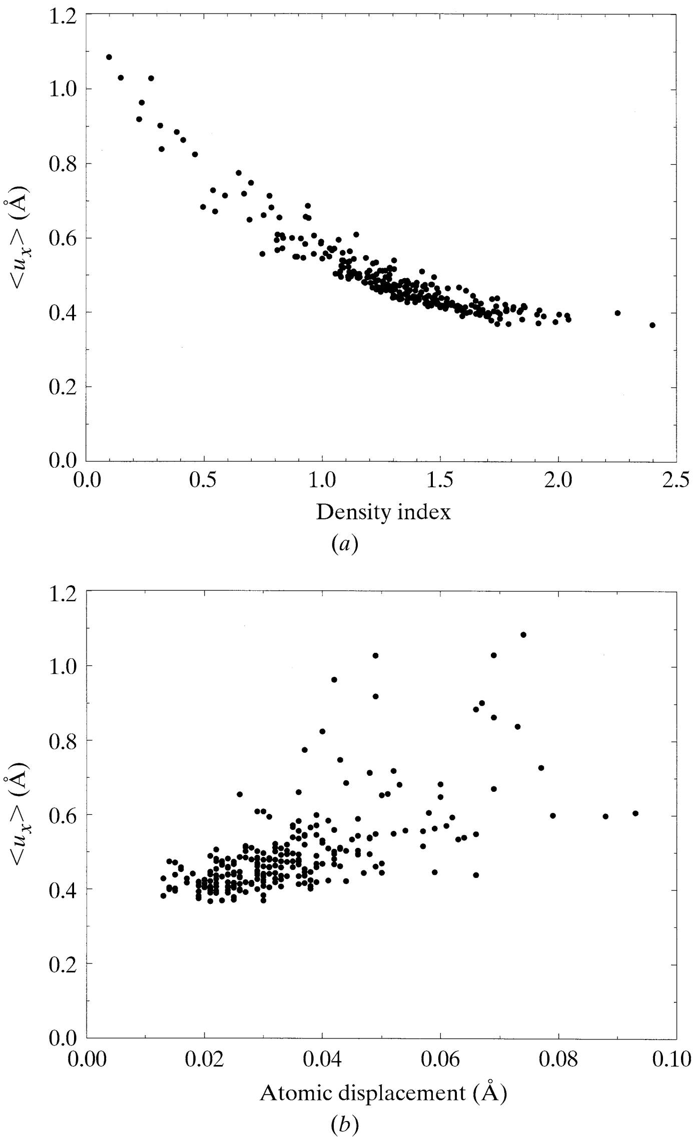


Figure 21.2.3.5

Pairwise correlations between the various local quality indicators computed by *SFCHECK*. (*a*) Correlation between the average residue *B* factor and the density index, and (*b*) between the *B* factor and the atomic displacement. The values displayed were computed for residues in the crystal structure of carboxypeptidase (1YME). The meaning of the parameters displayed is given in Table 21.2.3.3.

that the two measures are rather strongly correlated in individual structures. An indicator such as this one, and ultimately the atomic s.u.'s themselves, should be better suited for analysing and comparing the trends in the quality of specific regions of the model across different structures.

21.2.4. Atomic resolution structures

With improved techniques of crystallization and data collection using synchrotron radiation and cryogenic cooling, an increasing number of protein crystal structures are being determined at atomic resolution (1.2 Å or better). With atomic resolution data, refinement can be performed that requires much less strict compliance with prior knowledge of the expected geometry. Although some restraints must still be imposed, especially to deal with more flexible regions, and hence biases remain, it might be expected that these structures provide more precise information on the 'true' geometrical and stereochemical properties of proteins. Ultimately, one would want to re-derive these properties using only atomic resolution structures, but their number is at present too limited to provide sufficient data for a meaningful statistical analysis.

In the meantime, atomic resolution protein structures have been used to check geometric and conformational parameters that have been derived from other sources, including small-molecule crystals and the larger set of proteins determined at various levels of resolution (Longhi *et al.*, 1997; EU 3-D Validation Network, 1998). The EU 3-D Validation Network study showed that in the atomic resolution structures, most of the geometrical validation parameters are more tightly clustered about their mean value than in structures determined at lower resolution, including tighter clustering in the core regions of the Ramachandran plot, tighter clustering of the atomic volumes and smaller s.u.'s in the distributions of the χ_1, χ_2 dihedral angles. In contrast, the ω torsion angle about the peptide bond exhibits a wider distribution, with a mean of 179.0 (56)° compared to 179.6 (47)° previously computed in protein structures determined at various resolution levels (Morris *et al.*, 1992).

Recently, atomic resolution structures have also been used to derive atomic s.u values for proteins. Remarkably, the estimated coordinate errors for concanavalin A at 0.94 Å (Deacon *et al.*, 1997) were found to be equivalent to those of small-molecule crystal structures, despite the large size of the protein (237 residues).

Atomic resolution structures of proteins and other macromolecules thus promise to represent a valuable source of accurate information on geometric and conformational parameters of these molecules. But the analysis and validation of such structures also brings about additional complications, such as, for example, the problem of dealing with equilibria between multiple conformations, which atomic resolution data tend to resolve with much higher detail and accuracy. Handling these equilibria will require an adaptation of the current validation procedures.

21.2.5. Concluding remarks

The coming years will see an ever-increasing number of crystal structures of proteins and nucleic acids determined at high resolution and a substantial growth in the number of atomic resolution structures. This will most certainly help in obtaining better data on the geometric and stereochemical parameters of these macromolecules and thus improve the target values for both refinement and structure validation. It should also make it possible to derive better criteria for evaluating the agreement of the model with the electron density and to improve upon and generalize comprehensive and systematic approaches, such as that implemented in the software *SFCHECK*.

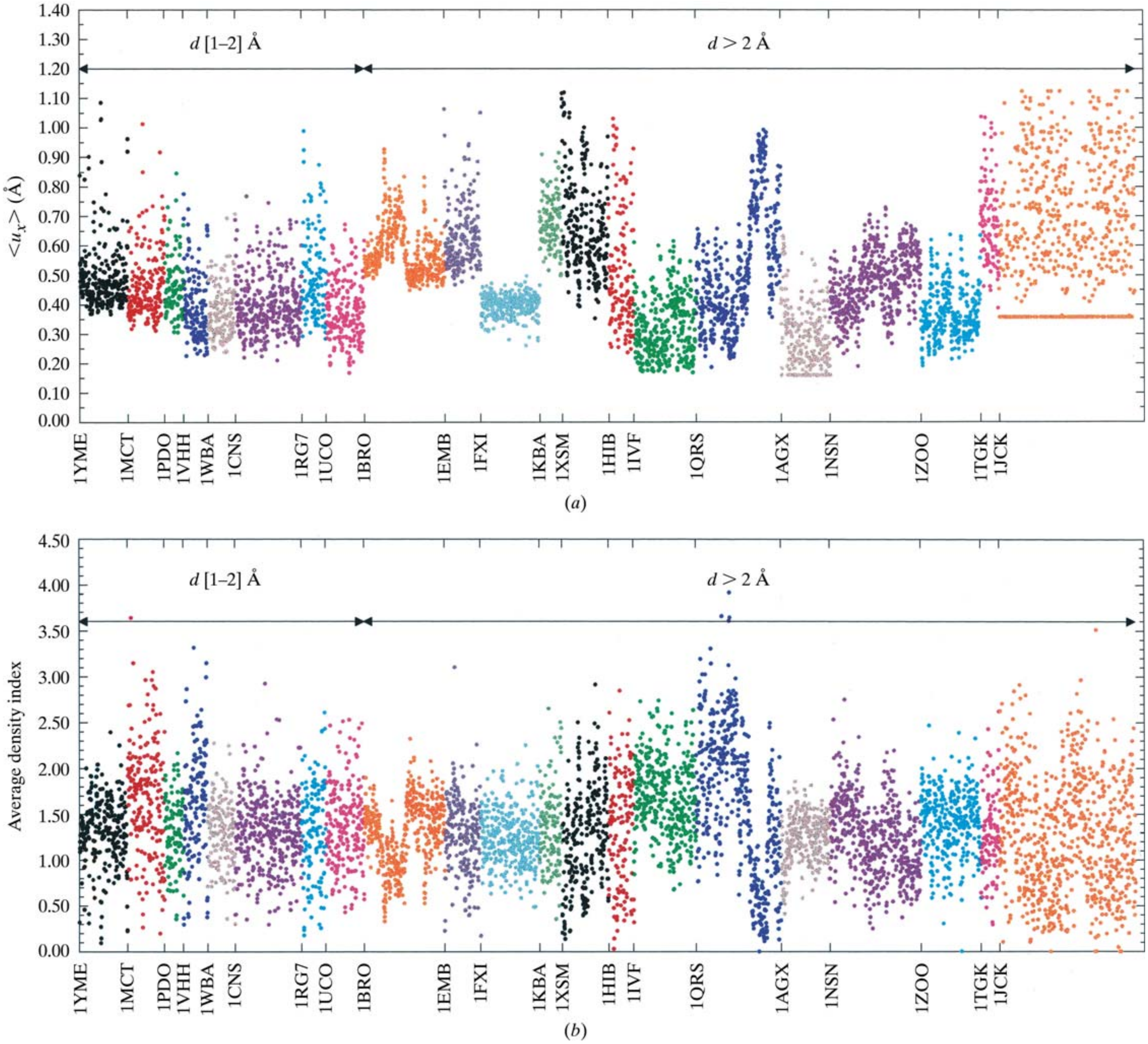


Figure 21.2.3.6

B factors and density indices for residues across different structures. (*a*) Average *B* values in residues of 21 protein structures; (*b*) average density indices of the same set of residues and structures. The 21 protein structures analysed are from the following PDB entries: 1YME, 1MCT, 1PDO, 1VHH, 1WBA, 1CNS, 1RG7, 1UCO, 1BRO, 1EMB, 1FXI, 1KBA, 1XSM, 1HIB, 1IVF, 1QRS, 1AGX, 1NSN, 1ZOO, 1TGK, 1JCK.

References

Abagyan, R. A. & Totrov, M. M. (1997). *Contact area difference (CAD): a robust measure to evaluate accuracy of protein models*. *J. Mol. Biol.* **268**, 678–685.

Agarwal, R. C. (1978). *A new least-squares refinement technique based on the fast Fourier transform algorithm. Acta Cryst.* A**34**, 791–809.

Alard, P. (1991). *Calcul de surface et d'énergie dans le domaine des macromolécules*. PhD thesis dissertation, Université Libre de Bruxelles, Belgium.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. Acta Cryst.* B**35**, 2331–2339.

Allen, F. H., Kennard, O. & Taylor, R. (1983). *Systematic analysis of structural data as a research technique in organic chemistry*. *Acc. Chem. Res.* **16**, 146–153.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J. Mol. Biol.* **112**, 535–542.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K., Westbrook, J. & Fitzgerald, P. M. D. (1997). *Macromolecular Crystallographic Information File*. *Methods Enzymol.* **277**, 571–590.

Brünger, A. T. (1992*a*). *X-PLOR manual.* Version 3.1. New Haven: Yale University Press.

Brünger, A. T. (1992*b*). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–474.

Brünger, A. T. (1997). *Free R value: cross-validation in crystallography. Methods Enzymol.* **277**, 366–396.

Chapman, M. S. (1995). *Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. Acta Cryst.* A**51**, 69–80.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: nitrogenous bases. J. Am. Chem. Soc.* **118**, 509–518.

Clowney, L., Westbrook, J. D. & Berman, H. M. (1999). *CIF applications. XI. A La Mode: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models. J. Appl. Cryst.* **32**, 125–133.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography. Acta Cryst.* D**50**, 760–763.

Cruickshank, D. W. J. (1949). *The accuracy of electron-density maps in X-ray analysis with special reference to dibenzyl. Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1965). *Errors in least-squares methods.* In *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 112–116. Oxford: Pergamon Press.

Cruickshank, D. W. J. (1996). *Protein precision re-examined: Luzzati plots do not estimate final errors.* In *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 11–22. Warrington: Daresbury Laboratory.

Das, U., Chen, S., Fuxreiter, M., Vaguine, A. A., Richelle, J., Berman, H. M. & Wodak, S. J. (2001). *Checking nucleic acid crystal structures. Acta Cryst.* D**57**, 813–828.

Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *The structure of concanavalin A and its bound solvent determined with small-molecule accuracy at 0.94 Å resolution. J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.

Dodson, E., Kleywegt, G. J. & Wilson, K. (1996). *Report of a workshop on the use of statistical validators in protein X-ray crystallography. Acta Cryst.* D**52**, 228–234.

Eisenberg, D., Luthy, R. & Bowie, J. U. (1997). *VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol.* **277**, 396–404.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

EU 3-D Validation Network (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J. Mol. Biol.* **276**, 417–436.

Finney, J. L. (1975). *Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J. Mol. Biol.* **96**, 721–732.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents. J. Am. Chem. Soc.* **118**, 519–529.

Gellatly, B. J. & Finney, J. L. (1982). *Calculation of protein volumes: an alternative to the Voronoi procedure. J. Mol. Biol.* **161**, 305–322.

Harata, K., Abe, Y. & Muraki, M. (1998). *Full-matrix least-squares refinement of lysozymes and analysis of anisotropic thermal motion. Proteins*, **30**, 232–243.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Volume changes on protein folding. Structure*, **2**, 611–649.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins*, **26**, 363–376.

Hooft, R. W. W., Sander, C. & Vriend, G. (1997). *Objectively judging the quality of a protein structure from a Ramachandran plot. Comput. Appl. Biosci.* **13**, 425–430.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures. Nature (London)*, **381**, 272.

Jones, T. A., Zou, J.-Y., Cowan, S. W., Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst.* A**47**, 110–119.

Kabsch, W. & Sander, C. (1983). *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers*, **22**, 2577–2637.

Kleywegt, G. J. & Brünger, A. T. (1996). *Checking your imagination: applications of the free R value. Structure*, **4**, 897–904.

Kleywegt, G. J. & Jones, T. A. (1996). *Phi/psi-chology: Ramachandran revisited. Structure*, **4**, 1395–1400.

Kleywegt, G. J. & Jones, T. A. (1998). *Databases in protein crystallography. Acta Cryst.* D**54**, 1119–1131.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1998). *Validation of protein models derived from experiment. Curr. Opin. Struct. Biol.* **8**, 631–639.

Longhi, S., Czjzek, M., Lamzin, V., Nicolas, A. & Cambillau, C. (1997). *Atomic resolution (1.0 Å) crystal structure of Fusarium solani cutinase: stereochemical analysis. J. Mol. Biol.* **268**, 779–799.

MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). *Knowledge-based validation of protein-structure coordinates derived by X-ray crystallography and NMR spectroscopy. Curr. Opin. Struct. Biol.* **4**, 731–737.

MacKerell, A. D. Jr, Bashford, D., Bellott, M., Dunbrack, R. L. Jr, Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E. III, Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M. (1998). *All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B*, **102**, 3586–3616.

Melo, F. & Feytmans, E. (1997). *Novel knowledge-based mean force potential at atomic level. J. Mol. Biol.* **267**, 207–222.

Melo, F. & Feytmans, E. (1998). *Assessing protein structures with a non-local atomic interaction energy. J. Mol. Biol.* **277**, 1141–1152.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Stereochemical quality of protein structure coordinates. Proteins Struct. Funct. Genet.* **12**, 345–364.

Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *The effect of overall anisotropic scaling in macromolecular refinement. Newsletter on Protein Crystallography*, pp. 37–42. Warrington: Daresbury Laboratory.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Refinement of macromolecular structures by the maximum-likelihood method. Acta Cryst.* D**53**, 240–255.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures. Acta Cryst.* D**52**, 57–64.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures. J. Mol. Biol.* **264**, 121–136.

Ratnaparkhi, G. S., Ramachandran, S., Udgaonkar, J. B. & Varadarajan, R. (1998). *Discrepancies between the NMR and X-ray structures of uncomplexed barstar: analysis suggests that packing densities of protein structures determined by NMR are unreliable. Biochemistry*, **37**, 6958–6966.

Richards, F. M. (1974). *The interpretation of protein structures: total volume, group volume distributions and packing density. J. Mol. Biol.* **82**, 1–4.

Richards, F. M. (1985). *Calculation of molecular volumes and areas for structures of known geometry. Methods Enzymol.* **115**, 440–464.

Sayle, R. A. & Milner-White, E. J. (1995). *RASMOL: biomolecular graphics for all. Trends Biochem. Sci.* **20**, 374–376.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Conformations of the sugar–phosphate backbone in helical DNA crystal structures. Biopolymers*, **42**, 113–124.

Sheldrick, G. M. & Schneider, T. R. (1997). *SHELXL: high-resolution refinement. Methods Enzymol.* **277**, 319–343.

Sheriff, S. & Hendrickson, W. A. (1987). *Description of overall anisotropy in diffraction from macromolecular crystals. Acta Cryst.* A**43**, 118–121.

Sippl, M. J. (1990). *Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J. Mol. Biol.* **213**, 859–883.

Sippl, M. J. (1993). *Recognition of errors in three-dimensional structures of proteins. Proteins*, **17**, 355–362.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γB- and βB2-crystallin. Acta Cryst.* D**54**, 243–252.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. Acta Cryst.* D**55**, 191–205.

Voronoi, G. F. (1908). *Nouvelles applications des paramètres continus à la théorie des formes quadratiques*. *J. Reine Angew. Math.* **134**, 198–287.

Walter, R. L., Ealick, S. E., Friedman, A. M., Blake, R. C. II, Proctor, P. & Shoham, M. (1996). *Multiple wavelength anomalous diffraction (MAD) crystal structure of rusticyanin: a highly oxidizing cupredoxin with extreme acid stability*. *J. Mol. Biol.* **263**, 730–749.

Wodak, S. J. & Rooman, M. (1993). *Generating and testing protein folds*. *Curr. Opin. Struct. Biol.* **3**, 247–259.

Zhou, G., Wang, J., Blanc, E. & Chapman, M. S. (1998). *Determination of the relative precision of atoms in macromolecular structure. Acta Cryst.* D**54**, 391–399.

International Tables for Crystallography (2012). Vol. F, Chapter 21.3, pp. 677–683.

Chapter 21.3. Detection of errors in protein models

O. Dym, D. Eisenberg and T. O. Yeates

21.3.1. Motivation and introduction

The discovery of major errors in several protein structural models determined by X-ray crystallography has focused attention on methods of detecting and minimizing such errors. There are several sources of error in the determination of a protein structure. Errors enter not only in the collection of the experimental data, but especially in their interpretation. Limited diffraction resolution and poor phases frequently lead to electron-density maps that are difficult to interpret. As a result, preliminary protein models built into ambiguous maps often contain errors of various types. The different types of errors can be arranged in decreasing order of severity, as follows: mistracing of the protein chain due to uncertainty in backbone connectivity, misalignment or misregistration of residues, and misplacement of side-chain and backbone atoms. It is critical to be able to identify these problematic regions of a model so they can be given special attention during the iterative process of model building and atomic refinement.

During atomic refinement, the atomic coordinates of the macromolecule are adjusted to minimize an error function of two terms. The first term contains the discrepancies between the observed diffraction data and structure factors calculated from the model. The second term describes the deviations from ideal geometry, such as deviations in bond lengths, bond angles, planarity and other specific features. When refinement is complete, the residual errors in the separate terms are reported, with the discrepancies in the diffraction data embodied in the R value. These error values are usually taken as the first indicators of structure quality.

Beyond criteria that are explicitly minimized during refinement, other structural properties may be devised and evaluated. Some properties that have been investigated include the distribution of non-polar and polar residues both on the surface and in the interior of the protein, and preferred environments for different atom types and residues. These measures use the empirical knowledge gathered in the Protein Data Bank (PDB) to assess how ‘normal’ or ‘abnormal’ a given model is. The measures are also useful in cases in which the experimental diffraction data are not available (*e.g.* when assessing structures already in the data bank). Several programs that validate protein structural models on the basis of various structural properties are available. Among them are *PROCHECK* (Laskowski *et al.*, 1993), *WHAT IF* (Vriend, 1990; Vriend & Sander, 1993), *ERRAT* (Colovos & Yeates, 1993), and *VERIFY3D* (Lüthy *et al.*, 1992; Bowie *et al.*, 1991). The various programs have the same objectives, but differ in many important respects. The approaches differ with regard to the scale of the analysis (*e.g.* atom-based *versus* amino-acid based), the level of detail in the program output, and the degree to which the evaluated properties are independent of the refinement function.

21.3.2. Separating evaluation from refinement

Any property that has been constrained or heavily restrained during refinement of the atomic model, and any property that has been closely monitored during rebuilding, cannot be used as the sole criterion to assess or ‘prove’ the quality of the model. The reason is that if the atomic model is adjusted to optimize a particular property, that property no longer gives an unbiased measure of model accuracy. For example, most refinement programs operate by adjusting atomic positions to minimize the difference between observed and calculated structure-factor amplitudes, known as the R factor or R value. Since the R value is the target of the optimization procedure, it does not provide an *independent* measure of quality. As a result, the ordinary R value can be misleading. A much more reliable measure is the free R value (Brünger, 1992), which is calculated from a randomly selected subset of the diffraction data that are excluded from the atomic refinement calculations. The importance of using the free R value to monitor refinement is now widely accepted.

Likewise, independent criteria must be employed to judge protein models themselves, aside from the diffraction data. Typical atomic refinement protocols tightly restrain the obvious stereochemical terms, such as bond lengths, angles and planarity. Therefore, low deviation from ideal geometry cannot be presented as proof of the quality of the structure. Independent criteria must be based on higher-level geometric considerations. Several programs that include such evaluations are described here.

Criteria that are useful for assessing the validity of protein models are those that are not directly restrained during the process of refinement. The following three properties of protein models are of this type: (1) the main-chain dihedral angles; (2) the non-bonded interactions of protein atoms with other protein atoms and with the solvent; and (3) the packing of atoms within the structure. Each of these properties of a proposed model can be compared for consistency with the same property observed in a database of trustworthy structures. To the extent that the property deviates from the values observed for the proteins of the database, the proposed model is suspect. Some of these properties can be computed for each segment of a protein or for local regions in three-dimensional (3D) space. In this way, inaccurate regions within a proposed model can be identified.

21.3.3. Algorithms for the detection of errors in protein models and the types of errors they detect

21.3.3.1. *PROCHECK*

The *PROCHECK* (Laskowski *et al.*, 1993) suite of programs compares the stereochemistry of a proposed protein model to stereochemical features of known structures. The program provides an assessment of the overall quality of the model by comparing the model with well refined structures of the same resolution, and also highlights regions that may need further adjustment. The output of *PROCHECK* comprises a number of plots, together with detailed residue-by-residue listings of secondary-structure assignment, non-bonded interactions between different pairs of residues, main-chain bond lengths and bond angles, and peptide-bond planarity.

Copyright © 2012 International Union of Crystallography

The program also displays main-chain dihedral angles (φ and ψ) as a two-dimensional Ramachandran (Ramachandran & Sasisekharan, 1968) plot. The Ramachandran plot classifies each residue in one of three categories: 'allowed' conformations; 'partially allowed' conformations, which give rise to modestly unfavourable repulsion between non-bonded atoms, and which might be overcome by attractive effects such as hydrogen bonds; and 'disallowed' interactions which give highly unfavourable non-bonded interatomic distances. The Ramachandran plot can identify unacceptable clusters of φ–ψ angles, revealing possible errors made during model building and refinement. As opposed to covalent bond angles and bond lengths, the main-chain dihedral angles are not usually restrained during X-ray refinement and therefore can be used to validate the structural model independently. In practice, the Ramachandran plot is one of the simplest, most sensitive tools for assessing the quality of a protein model.

The *PROCHECK* suite is generally useful for assessing the quality of protein structures in various stages of completion. The Ramachandran analysis is especially informative. However, it is possible, at least in principle, to devise an incorrect model with fully acceptable main-chain and side-chain stereochemistry, so other methods must also be used to assess protein models.

21.3.3.2. WHAT IF

The molecular modelling and drug design program *WHAT IF* (Vriend, 1990) performs a large number of geometrical checks, comparing a proposed protein model to a set of canonical distances and angles. These parameters include bond lengths and bond angles, side-chain planarity, torsion angles, interatomic distances, unusual backbone conformations and the Ramachandran plot. New additions (Vriend & Sander, 1993) include a 'quality factor', and a number of checks for clashes between symmetry-related molecules. Starting from the hypothesis that atom–atom interactions are the primary determinant of protein folding, the program tests a protein model for proper packing by calculating a contact quality index. Each contact is characterized by its fragment type (80 types from the 20 residues), the atom type and the three-dimensional location of the atom relative to the local frame of the fragment. Sets of database-derived distributions are compared with the actual distribution in the protein model being tested. A good agreement with the database distribution produces a high contact quality index. A low packing score can indicate any of: poor packing, misthreading of the sequence, bad crystal contacts, bad contacts with a co-factor, or proximity to a vacant active site. The contact analysis available in *WHAT IF* can be used as an independent quality indicator during crystallographic refinement, or during the process of protein modelling and design.

21.3.3.3. VERIFY3D

The program *VERIFY3D* (Lüthy *et al.*, 1992; Bowie *et al.*, 1991) measures the compatibility of a protein model with its own amino-acid sequence. Each residue position in the 3D model is characterized by its environment and is represented by a row of 20 numbers in a '3D profile'. These numbers are the statistical preferences, called 3D–1D scores, of each of the 20 amino acids for this environment. Environments of residues are defined by three parameters: the area of the residue buried in the protein and inaccessible to solvent, the fraction of the side-chain area that is covered by polar atoms (O and N), and the local secondary structure. The 3D profile score, S, for the compatibility of the sequence with the model is the sum, over all residue positions, of the 3D–1D scores for the amino-acid sequence of the protein. The compatibility of segments of the sequence with their 3D structures can be assessed by plotting, against sequence number, the average 3D–1D score in a window of 21 residues. The 3D profile method rests on the observation that soluble proteins bury many hydrophobic side chains and not many polar residues.

Three applications for 3D profiles exist. The first is to assess the validity of protein models (Lüthy *et al.*, 1992). For 3D protein models known to be correct, the 3D profile score, S, for the compatibility of the amino-acid sequence with the environments formed by the model is high. In contrast, S for the compatibility with its sequence of a totally or partially wrong 3D protein model is generally low. Therefore, models that are largely incorrect or models that contain a small number of improperly built segments can be detected by low-scoring regions in the 3D profile. However, not all faulty regions are always evident directly from the profile, particularly if the misbuilt regions are at the termini, where they are obscured by the windowing procedure. The second application is to assess which is the stable oligomeric state of the folded protein, by comparing the accessibility (buried or exposed) of amino-acid side chains in the monomeric and oligomeric state (Eisenberg *et al.*, 1992). The third application is to identify other protein sequences which are folded in the same general pattern as the structure from which the profile was prepared (Bowie *et al.*, 1991). Predicting a protein structure from sequence requires a link between 3D structure and 1D sequence. The program *VERIFY3D* provides this link by reducing a 3D structure to 1D string of environmental classes. Therefore the method can be used to evaluate any protein model or to measure the compatibility of any protein structure with its amino-acid sequence.

21.3.3.4. ERRAT

The program *ERRAT* (Colovos & Yeates, 1993) analyses the relative frequencies of noncovalent interactions between atoms of various types. It can be viewed as an extension of the earlier 3D profile approach from the residue level to the atom level. Three types of atoms are considered (C, N and O), and consequently six types of interactions are possible (CC, CN, CO, NN, NO and OO).

ERRAT operates under the hypothesis that different atom types will be distributed non-randomly with respect to each other in proteins due to complex geometric and energetic considerations, and that structural errors will lead to detectable anomalies in the pattern of interactions. Assessment of the non-bonded interactions is subject to the following restrictions: the distance between the two atoms in space is less than some preset limit, typically 3.5 Å, and the atoms within the same residue or those that are covalently bonded to each other are not considered. For each nine-residue segment of sequence, the non-bonded contacts to other atoms in the protein are tallied by atomic interaction type and the result is divided by the total number of interactions. This gives a list, or six-dimensional vector, of fractional interaction frequencies that add up to unity. In this way, each nine-residue fragment generates one point in a five-dimensional space; only five of the six fractional values are independent. A large number of observations were extracted from reliable high-resolution structures and used to establish a multivariate five-dimensional normal distribution for accurate protein structures. This distribution is used to evaluate the probability that a given

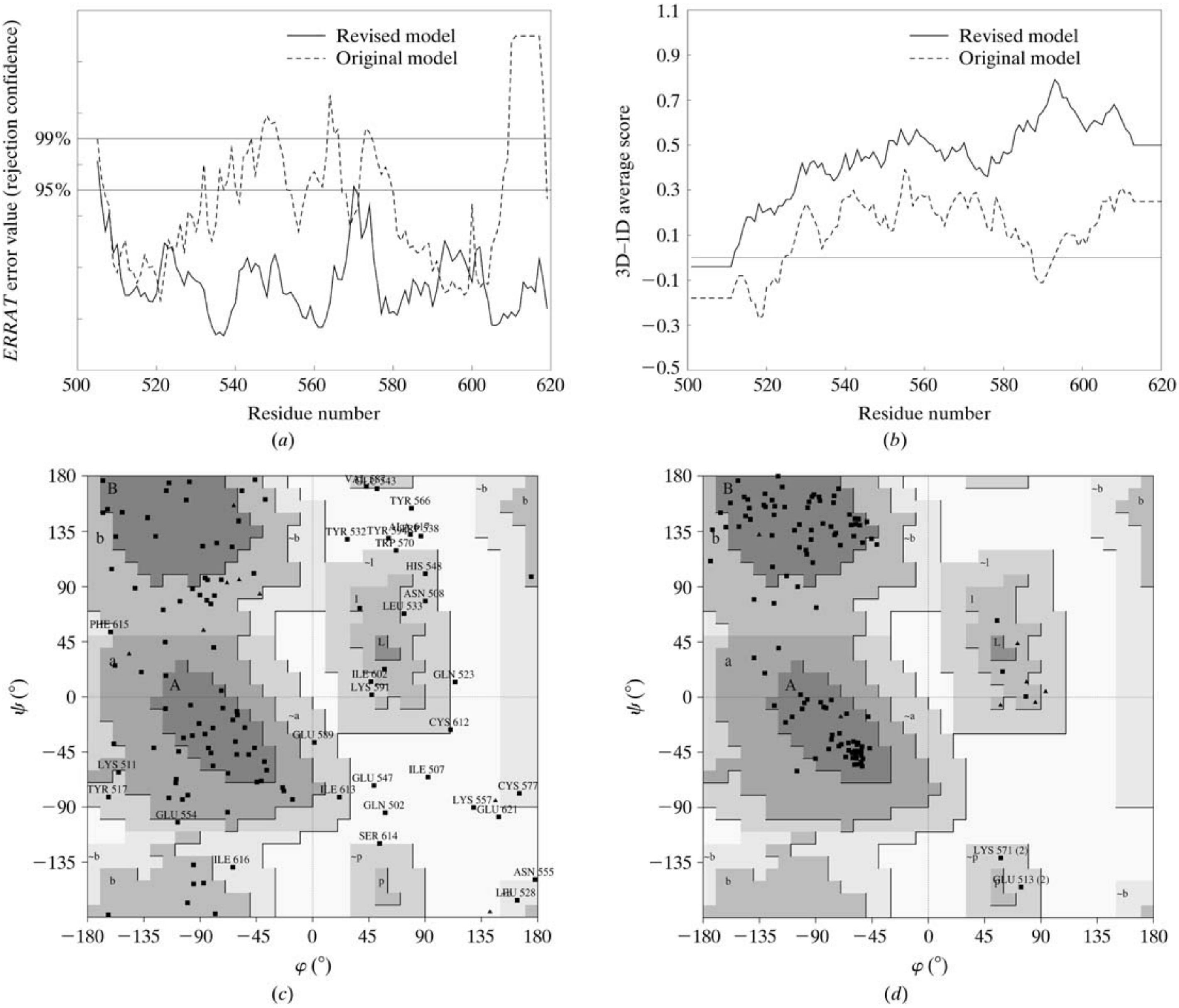


Figure 21.3.5.1

Detection of errors in the small subunit of ribulose-1,5-bisphospate carboxylase/oxygenase (RuBisCO). (*a*) *ERRAT* plot of the error function in a nine-residue sliding window, the centre of which is at the sequence position indicated by the horizontal axis. The solid bold line represents the revised structure and the dashed line the original structure. The thin solid lines indicate the 95% and 99% confidence limits for rejection. A region above the 95% line can be judged incorrect with 95% certainty. (*b*) *VERIFY3D* profile-window plots for the revised (bold) and original (dashed) models. The vertical axis gives the average 3D–1D score for residues within a 21-residue sliding window. Regions that score below zero are suspect. (*c*) Ramachandran diagram from *PROCHECK* of the initial structure of RuBisCO. The main-chain dihedral angle φ (N—Cα bond) is plotted *versus* ψ (Cα—C bond). All non-glycine residues outside the allowed regions are marked. (*d*) Ramachandran plot for the refined RuBisCO structure.

set of interactions from a protein model in question is correct. Since the *ERRAT* evaluation is based on a normal distribution calibrated on a reliable database, it is straightforward to estimate the likelihood that each region of a candidate protein model is incorrect. This method provides an unbiased and statistically sound tool for identifying incorrectly built regions in protein models.

21.3.4. Selection of database

Regardless of the specific approach or the specific criteria for validating structural models, a reliable reference database has to be chosen by careful selection of known structures. Suitable criteria to consider when selecting a database are: protein structures determined to resolutions of 2.5 Å or better, R factors less than 25%, and good geometry, particularly of the dihedral angles of the protein backbone. In addition, the database should include examples from many diverse classes of structures and at the same time avoid multiple identical structures.

21.3.5. Examples: detection of errors in structures

21.3.5.1. *Specific examples*

Several examples are presented of errors in structural models determined by X-ray crystallography that can be detected using validation methods. One is that of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), which was traced essentially backwards from a poor electron-density map (Chapman *et al.*, 1988). The program *ERRAT* finds that approximately 40% of the residues in this mistraced model are outside the 95% confidence limit (Fig. 21.3.5.1*a*). This limit is the error value above which a given region can be judged to be erroneous with 95% certainty, so a reliable model should exceed

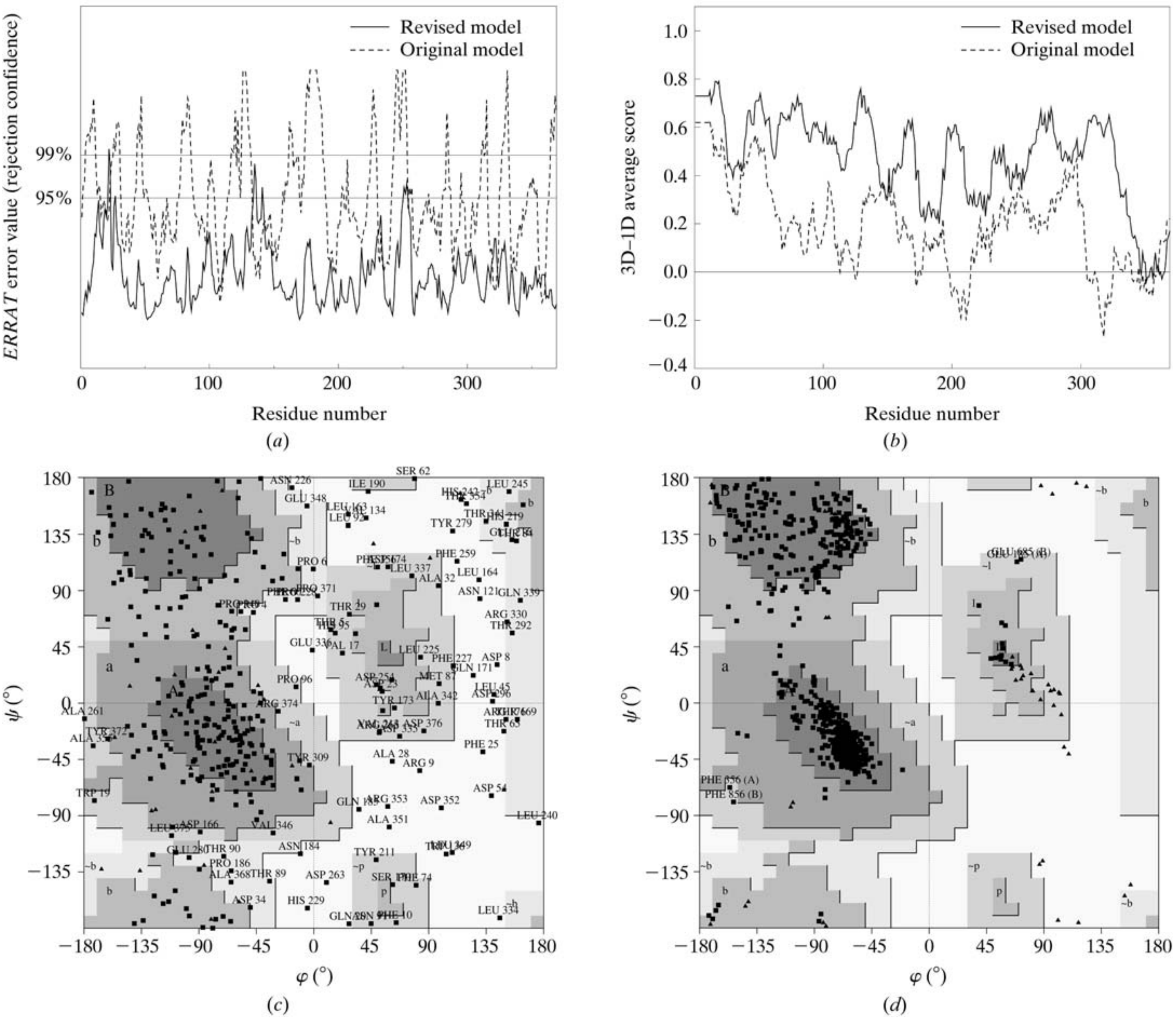


Figure 21.3.5.2
The detection of model errors due to refinement in an incorrect space group: an example (3xia.coor) from the archive of obsolete PDB entries. (*a*) *ERRAT* plot of the error function in a nine-residue sliding window. The solid bold line represents the revised structure and the dashed line represents the original structure. The thin solid lines indicate the 95% and 99% confidence limits for rejection. (*b*) *VERIFY3D* profile-window plots for the revised (bold) and original (dashed) models. The vertical axis gives the average 3D–1D score for residues within a 21-residue sliding window. (*c*) Ramachandran diagram of the original structure. All non-glycine residues outside the allowed regions are marked. (*d*) Ramachandran diagram for the revised structure.

this value over less than 5% of its length. The final model of RuBisCO (Curmi *et al.*, 1992) shows only 2% of the residues outside *ERRAT*'s 95% confidence limit. Similarly, the 3D profile calculated from *VERIFY3D* for the erroneous model (Fig. 21.3.5.1*b*) gives a total score of 15 when matched to the sequence of the small subunit of RuBisCO. This score is well below the expected value of 58 for the correct structure of this length. Indeed, the 3D profile of the correct model (Curmi *et al.*, 1992) (Fig. 21.3.5.1*b*) of RuBisCO has a score of 55. *PROCHECK* and *WHAT IF* also identify stereochemistry problems in the original model, including deviant bond angles and bond lengths, many residues in the disallowed Ramachandran regions (Fig. 21.3.5.1*c*), bad peptide-bond planarity, and bad non-bonded interactions. In contrast, most amino-acid residues of the correct RuBisCO model are in the allowed regions of the Ramachandran plot (Fig. 21.3.5.1*d*) with good overall geometry.

The archive of obsolete PDB entries maintained by the San Diego Supercomputer group (http://pdbobs.sdsc.edu) includes old versions of protein structures that have been withdrawn and/or replaced by the depositor with a newer version. One example is that of a protein (3xia.coor) originally solved to 3 Å in the wrong space group and later to 1.8 Å in the correct space group (1xya.coor). The *ERRAT* program reveals problems in the original model, with 45% of the residues outside the 95% confidence limit (Fig. 21.3.5.2*a*). The more recent model has only 1.5% of the residues outside the 95% confidence limit. The problem in the original model is also illustrated by the *VERIFY3D* plot (Fig. 21.3.5.2*b*) for which the average score is often below the value of 0.1 and dips below zero at four points. In contrast, the *VERIFY3D* plot of the revised model shows no dips below zero. Poor stereochemistry is also apparent in the Ramachandran plot of the original model (Fig. 21.3.5.2*c*). Only 38% of the backbone dihedral angles lie in the most favoured regions, compared to 93.8% in the revised model (Fig. 21.3.5.2*d*).

The potential usefulness of error-detecting programs during model building is suggested by stages in the crystal structure

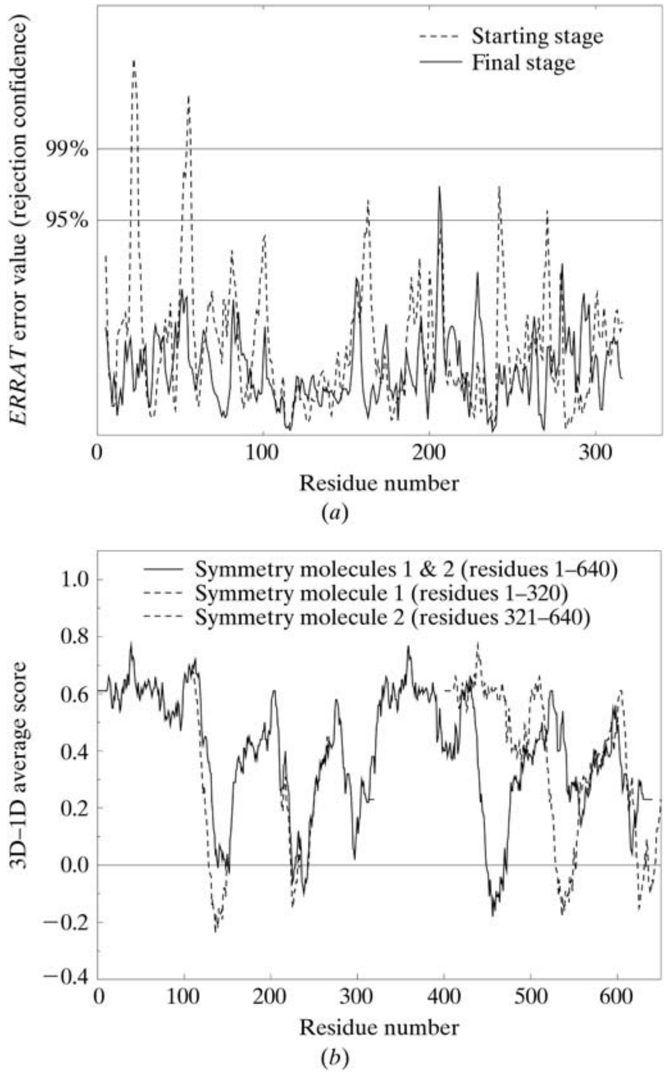


Figure 21.3.5.3
The usefulness of validation programs during model building is suggested by the example of the triacylglycerol lipase from *Pseudomonas cepacia* at different stages of atomic refinement. (*a*) Plot from *ERRAT* at the initial and final stages of refinement. (*b*) *VERIFY*3D profile-window plots of the final model. The dashed line represents symmetry molecule number 1 (residues 1–320) and symmetry molecule number 2 (residues 321–640) when not in contact with each other. The solid line represents symmetry molecule 1 and 2 when in contact. This plot illustrates that the state of oligomerization can affect the 3D profile plot, giving information on the oligomerization. See Bennett *et al.* (1994) for more details.

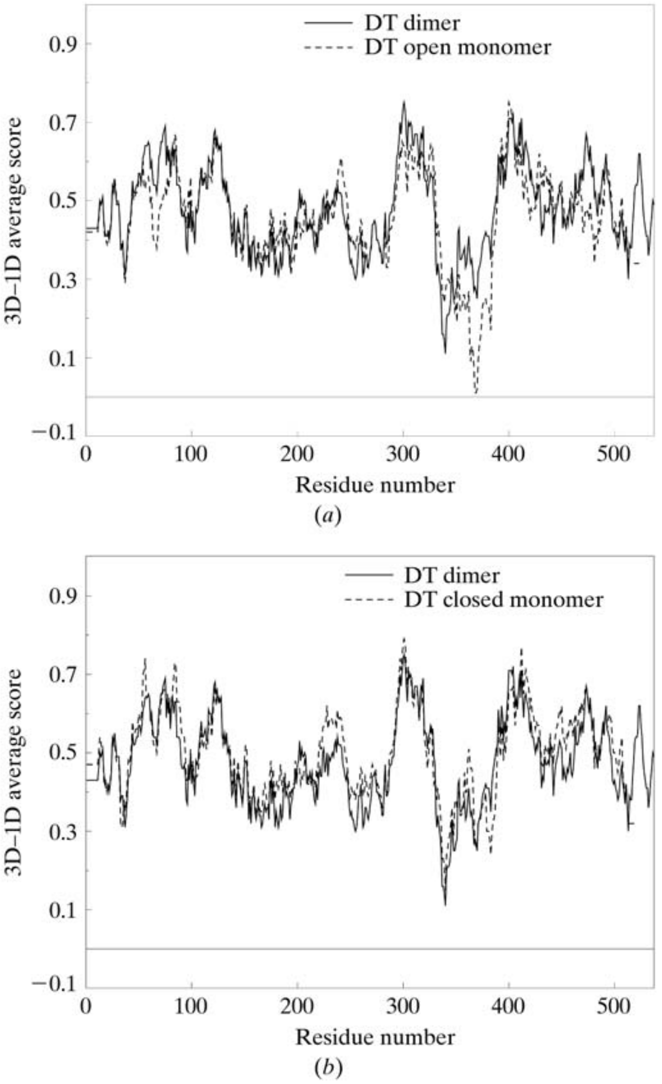


Figure 21.3.5.4
*VERIFY*3D profile plots of diphtheria toxin (DT) in three forms: open and closed monomers and the dimer. (*a*) DT open monomer (dashed), DT dimer (solid line). (*b*) DT closed monomer (dashed) and dimer (solid line). Notice that the hinge loop (residues 379–387) in the open monomer has a low profile score, and this structure is known to be unstable. The score is raised in the stable closed monomer and in the dimer.

determination of triacylglycerol lipase from *Pseudomonas cepacia* (Kim *et al.*, 1997), which was solved by MIR. The authors kindly provided us with ten different models (assigned as stage number 1–10) along the course of model building and refinement. Regions where Cα positions shifted between initial and final models correlated with regions where the error functions improved. For example, the program *ERRAT* points at specific regions (*e.g.* 18–35 and 135–165) originally assigned as poly-alanine. When at the next stage of refinement these were changed to the actual amino-acid sequence, these regions behaved normally (Fig. 21.3.5.3*a*). This illustrates that *ERRAT* is able to illuminate problem areas in a structure.

*VERIFY*3*D* is sensitive to unusual environments in proteins. An illustration is offered by the structures of lipases, with and without their inhibitors. There are two general conformations known as 'closed' and 'open'. In the so-called 'closed' structure, the catalytic triad is buried underneath a helical segment, called a 'lid' (Brzozowski *et al.*, 1991), so that hydrophobic residues tend to be buried as observed in a 'normal' 3D profile. In the 'open' conformation, the lipid binding site becomes accessible to the solvent, and hydrophobic surfaces (residues 140–150 and 230–250) are exposed by the movement of the 'lid'. These hydrophobic exposed regions are strikingly shown in the 3D profile of the 'open' structure (Fig. 21.3.5.3*b*), which clearly reveals the two problematic regions (140–150 and 230–250) with profile scores below zero. The exposed hydrophobic residues 140–150 from one symmetry model make van der Waals interactions with hydrophobic residues 230–250 from a symmetry-related molecule (Kim *et al.*, 1997). These interactions are revealed as higher scores in those regions when inspecting the 3D profiles of the two symmetry-related molecules.

Another example of unusual environment is that of diphtheria toxin (DT), which exists as a monomer as well as a dimer. Monomeric DT is a Y-shaped molecule with three domains known as catalytic (C), transmembrane (T) and receptor binding domain (R). Crystal structures have been determined for both the 'closed' monomeric form and for a domain-swapped dimeric form (Bennett *et al.*, 1994). Upon dimerization, a massive

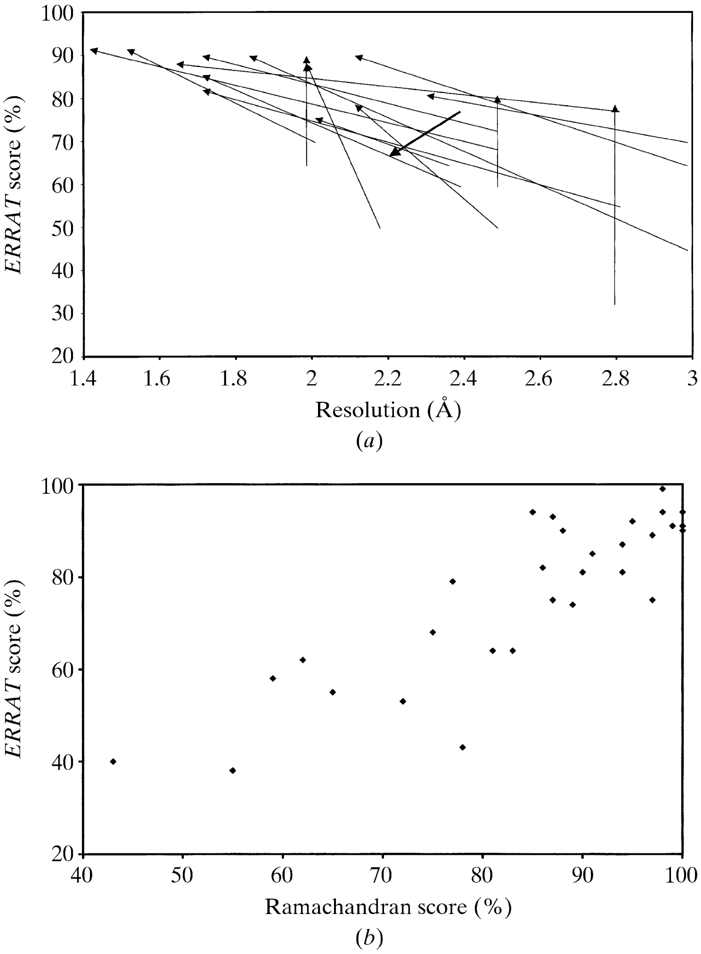


Figure 21.3.5.5
Evaluation of old and revised models in a database survey by *ERRAT*. (*a*) Percentage of residues within the 95% confidence limit given by *ERRAT* as a function of the resolution to which the structure was determined. Each arrow represents an obsolete structure (arrow tail) and the revised structure that replaced it (arrow head). The revised structures (typically analysed at finer resolution) show markedly improved *ERRAT* scores. (*b*) Correlation between the percentage of a structure within the 95% confidence limit according to *ERRAT*, and the percentage of residues in the most favoured regions of a Ramachandran diagram according to *PROCHECK*.

conformational rearrangement occurs and the entire R domain from each monomer of the dimer is interchanged with the other monomer. This involves breaking the noncovalent interactions between the R domain and the C and T domains and rotating the R domain by 180° with atomic movements up to 65 Å to produce the 'open' conformation. After rearrangement, each R domain reforms the same noncovalent interactions as it had in the monomer, but with the C and T domains of the other monomer. The existence of both open and closed forms of DT requires that large conformational changes occur in residues 379–387 (the hinge loop). The 3D profile of the 'open' form (Fig. 21.3.5.4*a*) shows low scores for these residues compared to the closed monomer or dimer (Fig. 21.3.5.4*b*). The higher scores of the open monomer are consistent with the greater stability of the monomer in the closed rather than the open conformation.

21.3.5.2. *Survey of old and revised structures*

The past two decades have seen a surge of development in the experimental techniques of crystal structure determination. As a consequence, many structures originally solved at low resolution were later determined at higher resolution, often starting with improved phases. The archive of obsolete PDB entries maintained by the San Diego Supercomputer group (http://pdbobs.sdsc.edu) served as a benchmark for evaluating the *ERRAT* program. For testing, 17 pairs of protein models were selected. Each pair comprised an obsolete entry and the revised model that replaced it. Using *ERRAT*, the overall quality of each model was expressed as a single number according to the fraction of the structure falling below the 95% confidence limit for rejection. The overall scores are significantly better for the revised structures, most of which were analysed at improved resolution (Fig. 21.3.5.5*a*). This result further demonstrates the utility of *ERRAT* for monitoring the model-building process. Furthermore, a strong correlation is found between the percentage of residues within the 95% confidence limit given by *ERRAT* and the percentage of residues in the most favoured regions of the Ramachandran plot of *PROCHECK* (Fig. 21.3.5.5*b*). In general, the problematic regions detected by the two programs agree with each other.

21.3.6. Summary

In order to ensure the quality of the growing protein structure databases, models must be evaluated carefully during and after the structure determination process. Model evaluation can incorporate two types of measures: agreement between the model and the experimental diffraction data, and agreement between the model and the database of known structures. The latter types use the atomic coordinates of the final model, but do not rely on the diffraction data. In recent years, powerful methods of this type have been developed.

The most informative and reliable model-evaluation criteria are those that measure properties not optimized as part of the automatic refinement procedure. The free *R* value has become important for monitoring the progress of atomic refinement for the same reason: it is based on reflections not included in refinement. We have focused here on two programs, *VERIFY3D* and *ERRAT*, which both evaluate high-level geometric properties not optimized during atomic refinement. Each offers the convenience of a single score over a sliding window along the protein sequence. Because *VERIFY3D* operates on the level of amino-acid residues, it is sensitive to errors on that scale, particularly those that affect the distribution of polar and nonpolar residues. *ERRAT* operates on the atomic level and has proven to be particularly useful for pinpointing local regions of protein models that require further adjustments. When used in combination, these methods and others can help crystallographers produce more accurate structural models of proteins.

21.3.7. Availability of software

The programs *ERRAT* and *VERIFY3D* are available on the World Wide Web for non-commercial applications. The URL for *VERIFY3D* is http://nihserver.mbi.ucla.edu/Verify_3D/ and the URL for *ERRAT* is http://nihserver.mbi.ucla.edu/ERRATv2/. *VERIFY3D* and *ERRAT* expect a coordinate file in PDB format. The programs return plots of the type shown in this chapter.

We thank Dr Kyeong Kyu Kim and Dr Se Won Suh of the Department of Chemistry, Seoul National University, Korea, for the models of triacylglycerol lipase. This work was supported by grants NIH GM 31299, DOE DE-FG03-87ER60615 and NSF MCB 9420769.

References

Bennett, M. J., Choe, S. & Eisenberg, D. (1994). *Domain swapping: entangling alliances between proteins*. *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). *A method to identify protein sequences that fold into a known three-dimensional structure*. *Science*, **253**, 164–170.

Brünger, A. T. (1992). *Free R value: a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Brzozowski, A. M., Derewenda, U., Derewenda, Z. S., Dodson, G. G., Lawson, D. M., Turkenburg, J. P., Bjorkling, F., Huge-Jensen, B., Patkar, S. A. & Thim, L. (1991). *A model for interfacial activation in lipases from the structure of fungal lipase-inhibitor complex*. *Nature (London)*, **351**, 491–494.

Chapman, M. S., Suh, S. W., Curmi, P. M., Cascio, D., Smith, W. W. & Eisenberg, D. (1988). *Tertiary structure of plant RuBisCO: domains and their contacts*. *Science*, **241**, 71–74.

Colovos, C. & Yeates, T. O. (1993). *Verification of protein structures: patterns of nonbonded atomic interactions*. *Protein Sci.* **2**, 1511–1519.

Curmi, P. M. G., Cascio, D., Sweet, R. M., Eisenberg, D. & Schreuder, H. (1992). *Crystal structure of the unactivated form of ribulose-1,5 bisphosphate carboxylase/oxygenase from tobacco refined at 2.0 Å resolution*. *J. Biol. Chem.* **267**, 16980–16989.

Eisenberg, D., Bowie, J. U., Lüthy, R. & Choe, S. (1992). *Three-dimensional profiles for analyzing protein sequence-structure relationships*. *Faraday Discuss. Chem. Soc.* **93**, 25–34.

Kim, K. K., Song, H. K., Shin, D. H., Hwang, K. Y. & Suh, S. W. (1997). *The crystal structure of a triacylglycerol lipase from Pseudomonas cepacia reveals a highly open conformation in the absence of a bound inhibitor*. *Structure*, **5**, 173–185.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures*. *J. Appl. Cryst.* **26**, 283–291.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). *Assessment of protein models with three-dimensional profiles*. *Nature (London)*, **356**, 83–85.

Ramachandran, G. N. & Sasisekharan, V. (1968). *Conformation of polypeptides and proteins*. *Adv. Protein Chem.* **23**, 283–438.

Vriend, G. (1990). *WHAT IF: a molecular modeling and drug design program*. *J. Mol. Graphics*, **8**, 52–56.

Vriend, G. & Sander, C. (1993). *Quality control of protein models: directional atomic contact analysis*. *J. Appl. Cryst.* **26**, 47–60.

International Tables for Crystallography (2012). Vol. F, Chapter 21.4, pp. 684–687.

Chapter 21.4. *PROCHECK*: validation of protein-structure coordinates

R. A. Laskowski, M. W. MacArthur and J. M. Thornton

21.4.1. Introduction

As in all scientific measurements, the parameters that result from a macromolecular structure determination by X-ray crystallography (*e.g.* atomic coordinates and B factors) will have associated uncertainties. These arise not only from systematic and random errors in the experimental data but also in the interpretation of those data. Currently, the uncertainties cannot easily be estimated for macromolecular structures due to the computer- and memory-intensive nature of the calculations required (Tickle *et al.*, 1998). Thus, more indirect methods are necessary to assess the reliability of different parts of the model, as well as the reliability of the model as a whole. Among these methods are those that rely on checking only the stereochemical and geometrical properties of the model itself, without reference to the experimental data (MacArthur *et al.*, 1994; Laskowski *et al.*, 1998). Here we describe *PROCHECK* (Laskowski *et al.*, 1993), which is one of these structure-validation methods.

The *PROCHECK* program computes a number of stereochemical parameters for the given protein model and compares them with 'ideal' values obtained from a database of well refined high-resolution protein structures in the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The results of these checks are output in easy-to-understand coloured plots in PostScript format (Adobe Systems Inc., 1985). Significant deviations from the derived standards of normality are highlighted as being 'unusual'.

The program's primary use is during the refinement of a protein structure; the highlighted regions can direct the crystallographer to parts of the structure that may have problems and which may need attention. It should be noted that outliers may just be outliers; they are not necessarily errors. Unusual features may have a reasonable explanation, such as distortions due to ligand binding in the protein's active site. However, if there are many oddities throughout the model, this could signify that there is something wrong with it as a whole. Conversely, if a model has good stereochemistry, this alone is not proof that it is a good model of the protein structure.

Because the program requires only the three-dimensional atomic coordinates of the structure, it can check the overall 'quality' of any model structure: whether derived experimentally by crystallography or NMR, or built by homology modelling. In the case of NMR-derived structures, it is useful to compare the protein geometry across the whole ensemble. An extended version of *PROCHECK*, called *PROCHECK-NMR*, is available for this purpose (Laskowski *et al.*, 1996), but will not be described here.

Note that *PROCHECK* only examines the geometrical properties of protein molecules; it ignores DNA/RNA and other non-protein molecules in the structure, except in so far as checking that the non-bonded contacts these make with the protein do not violate a fixed distance criterion.

21.4.2. The program

PROCHECK is in fact a suite of separate Fortran and C programs which are run successively *via* a shell script. The programs first 'clean up' the input PDB file, relabelling certain side-chain atoms according to the IUPAC naming conventions (IUPAC–IUB Commission on Biochemical Nomenclature, 1970), then calculate all the protein's stereochemical parameters to compare them against the norms, and finally generate the PostScript output and a detailed residue-by-residue listing. Hydrogen and atoms with zero occupancy are omitted from the analyses and, where atoms are found in alternate conformations, only the highest-occupancy conformation is retained.

The source code for all the programs is available at http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html. It has also been incorporated into the *CCP4* suite of programs (Collaborative Computational Project, Number 4, 1994) at http://www.dl.ac.uk/CCP/CCP4/main.html, and can be run directly *via* the web from the Biotech Validation Server at http://biotech.embl-ebi.ac.uk:8400/.

21.4.3. The parameters

Table 21.4.3.1 shows the principal stereochemical parameters used by *PROCHECK*, based on the analysis of Morris *et al.* (1992), who looked for measures that are good indicators of protein quality. The table shows the original parameters together with a more up-to-date set derived from a more recent data set including a number of atomic resolution structures (*i.e.* those solved to 1.4 Å resolution or better).

Table 21.4.3.1
Summary of expected values for stereochemical parameters in well resolved structures

Parameter	Old	New
% φ, ψ in core	>90.0%	>90.0%
χ_1 *gauche*$^-$	+64.1 ± 15.7°	+63.2 ± 11.4°
χ_1 *trans*	+183.6 ± 16.8°	+182.7 ± 13.1°
χ_1 *gauche*$^+$	−66.7 ± 15.0°	−66.0 ± 11.2°
χ_1 pooled standard deviation	±15.7°	±11.8°
χ_2 *trans*	+177.4 ± 18.5°	+177.2 ± 15.1°
χ_3 S—S bridge (left-handed)	−85.8 ± 10.7°	−84.8 ± 8.5°
χ_3 S—S bridge (right-handed)	+96.8 ± 14.8°	+92.2 ± 10.8°
Proline φ	−65.4 ± 11.2°	−64.6 ± 10.2°
α-Helix φ	−65.3 ± 11.9°	−65.5 ± 11.1°
α-Helix ψ	−39.4 ± 11.3°	−39.0 ± 9.8°
ω *trans*	+179.6 ± 4.7°	+179.5 + 6.0°
Cα—N—C$'$—Cβ (ζ) virtual torsion angle	+33.9 ± 3.5°	+34.2 ± 2.6°

For the most part, the parameters given in Table 21.4.3.1 are not included in standard refinement procedures and so are less likely to be biased by them. They can thus provide a largely independent and unbiased validation check on the geometry of each residue and hence point to regions of the protein structure that are genuinely unusual.

As more atomic resolution structures become available (Dauter *et al.*, 1997), these parameters will be improved. Because of their high data-to-parameter ratio, such structures can be refined using less strict restraints, and hence contain a smaller degree of bias in their geometrical properties – at least for the well ordered parts of the model. Such information moves us a step closer to an understanding of the 'true' geometrical and

Copyright © 2012 International Union of Crystallography

conformational properties of proteins in general and, one day, the target parameters will be derived exclusively from such structures.

PROCHECK also checks main-chain bond lengths and bond angles against the 'ideal' values given by the Engh & Huber (1991) analysis of small-molecule structures in the Cambridge Structural Database (CSD) (Allen *et al.*, 1979). Unlike the above parameters, these geometrical properties are usually restrained during refinement, and, furthermore, the Engh & Huber (1991) targets are the ones most commonly applied. Thus analyses of these values merely reflect the refinement protocol used and do not provide meaningful indicators of local or overall errors. However, the plots clearly show any wayward outliers which can nevertheless indicate problem regions in the structure.

21.4.4. Which parameters are best?

Possibly the most telling and useful of the 'quality' indicators for a protein model is the Ramachandran plot of residue φ–ψ torsion angles. This can often detect gross errors in the structure (Kleywegt & Jones, 1996*a*,*b*). In the original Ramachandran plot (Ramachandran *et al.*, 1963; Ramakrishnan & Ramachandran, 1965), the 'allowed' regions were defined on the basis of simulations of dipeptides. In the *PROCHECK* version, the different regions of the plot are defined on the basis of how densely they are populated with data points taken from a database of well refined protein structures. The regions are: core, allowed, generously allowed and disallowed.

The 'core' regions are particularly important; the points on the plot tend to converge towards these regions, and to cluster more tightly within them, as one goes from structures solved at low resolution to those solved at high resolution (Morris *et al.*, 1992). This trend has recently been confirmed by Wilson *et al.* (1998), who looked at the case of atomic resolution structures. It has also been analysed in terms of 'attractors' at the most favourable regions of the plot; as the resolution improves, so the points are drawn towards these attractors (Walther & Cohen, 1999).

Fig. 21.4.4.1 shows the original *PROCHECK* Ramachandran plot and a more up-to-date version. The original was based on all 462 structures known at that time (1989/90), while the more recent one, generated in 1998, is based on 1128 non-identical (*i.e.* having a sequence identity <95%) structures. It can be seen that the second plot has core regions which are much tighter than the original, and this is primarily due to the increase in the number of very high resolution structures giving a more accurate representation of the tight clustering in the most favourable regions.

Another parameter that seems to be a particularly sensitive measure of quality is the standard uncertainty (s.u.) of the χ torsion angles. Morris *et al.* (1992) found that the average values of a protein's χ_1 and χ_2 torsion angles are well correlated with the resolution at which the protein structure was solved. Although the data set was a fairly small one, the conclusion was borne out when tested on a larger set of more recent structures, including some solved to atomic resolution (Wilson *et al.*, 1998). This measure, however, cannot be relied on where side-chain conformations are either restrained or heavily influenced by the use of rotamer libraries.

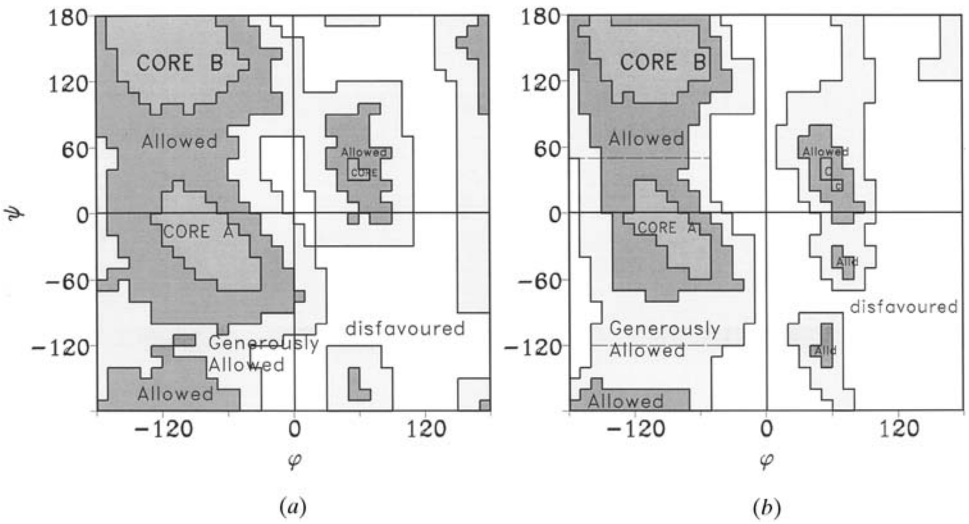
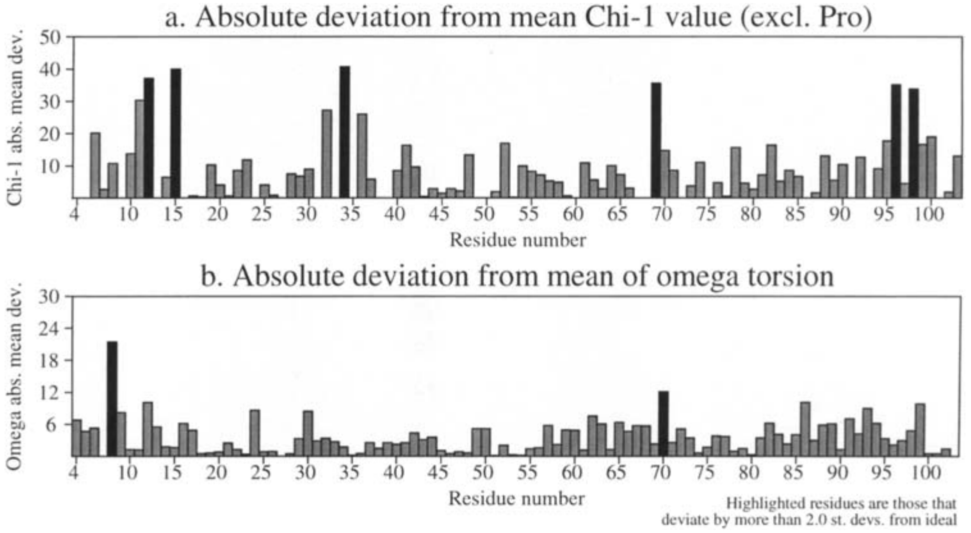


Figure 21.4.4.1
PROCHECK Ramachandran plots showing the different regions, shaded according to how 'favourable' the φ–ψ combinations are, for (*a*) the original version of the program (1992) and (*b*) an updated version based on a more recent data set (1998) including more high-resolution structures. The 'core' and other favourable regions of the plot are more tightly compressed in the new version, with the white, disfavoured regions occupying more of the space.


a. Absolute deviation from mean Chi-1 value (excl. Pro)

b. Absolute deviation from mean of omega torsion

Highlighted residues are those that deviate by more than 2.0 st. devs. from ideal

Figure 21.4.6.1
Two of the residue-property plots generated by *PROCHECK*. The plots shown here are (*a*) the absolute deviation from the mean of the χ_1 torsion angle (excluding prolines) and (*b*) the absolute deviation from the mean of the ω torsion angle. Usually, three such plots are shown per page and can be selected from a set of 14 possible plots. On each graph, unusual values (usually those more than 2.0 standard deviations away from the 'ideal' mean value) are highlighted.

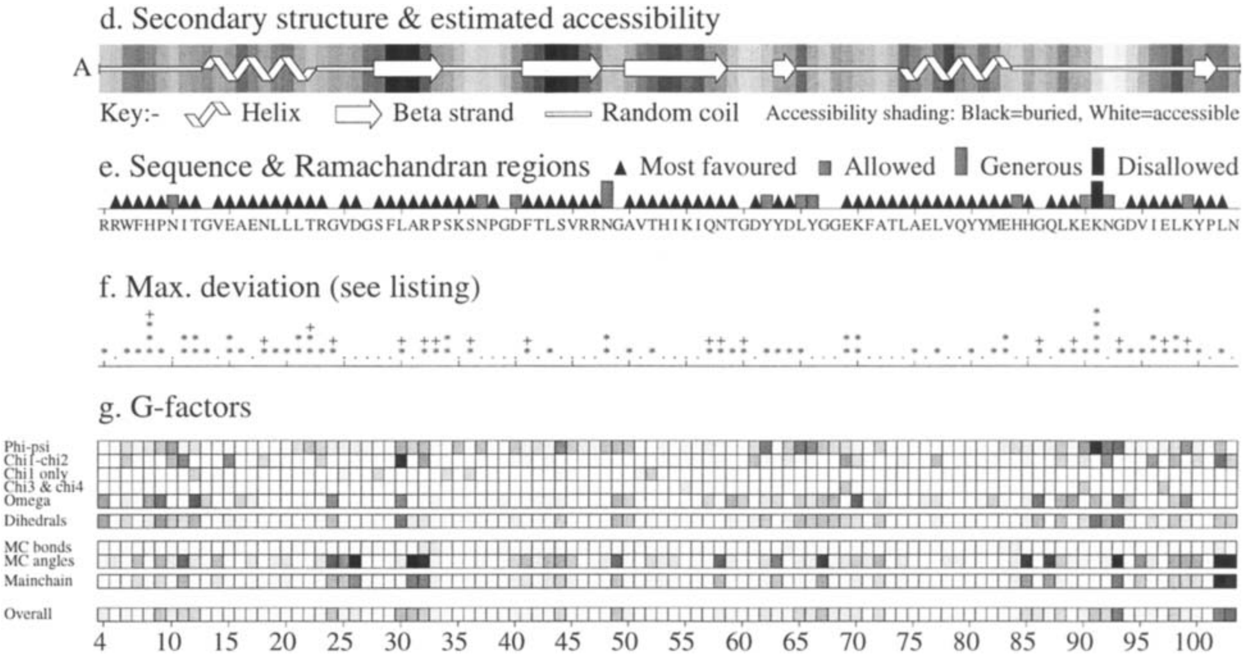


Figure 21.4.6.2

Schematic plots of various residue-by-residue properties, showing (*d*) the protein secondary structure, with the shading behind it giving an approximation to each residue's accessibility, the darker the shading the more buried the residue; (*e*) the protein sequence plus markers identifying the region of the Ramachandran plot in which the residue is located; (*f*) a histogram of asterisks and plus signs showing each residue's maximum deviation from one of the ideal values, as shown on the residue-by-residue listing; and (*g*) the residue '*G* factor' values for various properties, where the darker the square the more 'unusual' the property.

21.4.5. Input

The primary input to *PROCHECK* is the file containing the 3D coordinates of the protein structure to be processed. The file is required to be in PDB format. An additional input file is the parameter file that governs which plots are to be generated and deals with certain aspects of their appearance.

21.4.6. Output produced

The output of the program consists of a number of PostScript plots, together with a full listing of the individual parameter values for each residue, with any unusual geometrical properties highlighted. The listing also provides summaries for the protein as a whole. Figs. 21.4.6.1 and 21.4.6.2 show parts of one of the PostScript plots generated, showing the variation of various residue properties along the length of the protein chain. Unusual regions, which are highlighted on these plots, may require further investigation by the crystallographer.

21.4.7. Other validation tools

PROCHECK is merely one of a number of validation tools that are freely available, some of which are mentioned elsewhere in this volume. The best known are *WHATCHECK* (Hooft *et al.*, 1996), *PROVE* (Pontius *et al.*, 1996), *SQUID* (Oldfield, 1992) and *VERIFY3D* (Eisenberg *et al.*, 1997). Tools such as *OOPS* (Kleywegt & Jones, 1996*b*) or the X-build validation in *QUANTA* (MSI, 1997) provide standard tests on the geometry of a structure and provide lists of residues with unexpected features, which make it easy to check electron-density maps at suspect points.

Significant contributors to the programs in the *PROCHECK* suite include David K. Smith, E. Gail Hutchinson, David T. Jones, J. Antoon C. Rullmann, A. Louise Morris and Dorica Naylor. Part of the development work was funded by a grant from the EU Framework IV Biotechnology programme, contract CT96–0189.

References

Adobe Systems Inc. (1985). *PostScript Language Reference Manual*. Reading, MA: Addison-Wesley.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J. Mol. Biol.* **112**, 535–542.

Collaborative Computational Project, Number 4 (1994). *The CCP4 suite: programs for protein crystallography*. *Acta Cryst.* D**50**, 760–763.

Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1997). *The benefits of atomic resolution*. *Curr. Opin. Struct. Biol.* **7**, 681–688.

Eisenberg, D., Lüthy, R. & Bowie, J. U. (1997). *VERIFY3D: assessment of protein models with three-dimensional profiles*. *Methods Enzymol.* **277**, 396–404.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement. Acta Cryst.* A**47**, 392–400.

Hooft, R. W. W., Sander, C., Vriend, G. & Abola, E. E. (1996). *Errors in protein structures*. *Nature (London)*, **381**, 272.

IUPAC–IUB Commission on Biochemical Nomenclature (1970). *Abbreviations and symbols for the description of the conformation of polypeptide chains*. *J. Mol. Biol.* **52**, 1–17.

Kleywegt, G. J. & Jones, T. A. (1996*a*). *Phi/psi-chology: Ramachandran revisited*. *Structure*, **4**, 1395–1400.

Kleywegt, G. J. & Jones, T. A. (1996*b*). *Efficient rebuilding of protein structures*. *Acta Cryst.* D**52**, 829–832.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1998). *Validation of protein models derived from experiment*. *Curr. Opin. Struct. Biol.* **8**, 631–639.

Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). *AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J. Biomol. Nucl. Magn. Reson.* **8**, 477–486.

MacArthur, M. W., Laskowski, R. A. & Thornton, J. M. (1994). *Knowledge-based validation of protein structure coordinates derived by X-ray crystallography and NMR spectroscopy*. *Curr. Opin. Struct. Biol.* **4**, 731–737.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Stereochemical quality of protein structure coordinates*. *Proteins*, **12**, 345–364.

MSI (1997). *QUANTA*. MSI, 9685 Scranton Road, San Diego, CA 92121–3752, USA.

Oldfield, T. J. (1992). *SQUID: a program for the analysis and display of data from crystallography and molecular dynamics*. *J. Mol. Graphics*, **10**, 247–252.

Pontius, J., Richelle, J. & Wodak, S. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures*. *J. Mol. Biol.* **264**, 121–136.

Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *Stereochemistry of polypeptide chain configurations*. *J. Mol. Biol.* **7**, 95–99.

Ramakrishnan, C. & Ramachandran, G. N. (1965). *Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units*. *Biophys. J.* **5**, 909–933.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γB- and γB2-crystallin*. *Acta Cryst.* D**54**, 243–252.

Walther, D. & Cohen, F. E. (1999). *Conformational attractors on the Ramachandran map*. *Acta Cryst.* D**55**, 506–517.

Wilson, K. S., Butterworth, S., Dauter, Z., Lamzin, V. S., Walsh, M., Wodak, S., Pontius, J., Richelle, J., Vaguine, A., Sander, C., Hooft, R. W. W., Vriend, G., Thornton, J. M., Laskowski, R. A., MacArthur, M. W., Dodson, E. J., Murshudov, G., Oldfield, T. J., Kaptein, R. & Rullmann, J. A. C. (1998). *Who checks the checkers? Four validation tools applied to eight atomic resolution structures*. *J. Mol. Biol.* **276**, 417–436.

International Tables for Crystallography (2012). Vol. F, Chapter 21.5, pp. 688–693.

Chapter 21.5. *KiNG* and kinemages

V. B. Chen, J. S. Richardson and D. C. Richardson

21.5.1. Introduction to aims and concepts

KiNG and the kinemages it displays provide molecular graphics, organized in an unusual way, that are of interest to crystallographers for uses that range from interactive illustrations for teaching and web display, to high-dimensional data display, to a dot-surface representation of all-atom van der Waals contacts, updated in real time by *Probe* to help guide model-to-map fitting.

A kinemage ('kinetic image') is an authored interactive three-dimensional illustration that allows open-ended exploration but has viewpoint, explanation and emphasis built in. A kinemage is stored as a human-readable flat ASCII text file that embodies the data structure and three-dimensional plotting information chosen by its author or user. The original kinemage viewer, *Mage* (Richardson & Richardson, 1992, 1994), is a pure graphics display program designed to show and edit kinemages, while *Prekin* constructs molecular kinemages from PDB (Protein Data Bank; Berman *et al.*, 2000) files. A spiritual successor to *Mage*, *KiNG* (Kinemage, Next Generation) is a Java-based kinemage reader (Chen *et al.*, 2009) and has evolved into a fully featured alternative to *Mage*. *KiNG* includes a Java program *Molikin*, based heavily on *Prekin*, to construct kinemages from PDB files or mmCIF files. The latest versions of *KiNG*, *Mage* and *Prekin* are available free and open source for Macintosh, PC or Linux from the kinemage web site (http://kinemage.biochem.duke.edu).

KiNG and *Mage* operate quite equivalently on the different platforms, and their graphical interfaces differ very little except where each has some unique feature for construction, drawing or modelling. All the major features of *KiNG* described here are also available in *Mage*, except for the 'backrub' tool (Davis *et al.*, 2006) and the web online display. *KiNG* has more convenient tools for superposition, electron-density display and built-in kinemage construction, while *Mage* has a number of historical tools for highly specialized purposes, such as a system for defining separate groups with rotatable bonds (such as an RNA suite or a ligand), and the ability to time, score and save exercise or test results. By policy, later versions of *Mage* can display all older kinemages back to 1992, and *KiNG* comes very close. This means that the kinemage format has standard forms but also recognizes many alternative forms likely to occur in human-edited files or those produced by other software (such as 'gray' *versus* 'grey', and commas, spaces or line breaks as field separators).

By design, *KiNG* has no internal knowledge of molecular structure (except for a few built-in libraries used for mutations and rebuilding). A collaboration between the author and the authoring program (*e.g. Molikin*) builds data organization into the kinemage itself. This two-layer approach has great advantages in flexibility, since an author can show things the programmer never imagined, including non-molecular three-dimensional relationships. Overall, just viewing kinemages demands less work and less expertise from the reader than most molecular graphics, but that ease of use depends on the authoring effort involved in making thoughtful choices, aided by the extensive on-screen editing capabilities described below.

Mage and *KiNG* are both designed to optimize visual comprehension: the understanding and communication of specific three-dimensional relationships inside complex molecules. Display speed has been given priority, to ensure good depth perception from smooth real-time rotation, even for a whole ribosome. The interface is extremely simple and transparent, and the colour palette has been carefully tuned for visual comparisons, contrasts and depth cueing. Immediate identification and measurement are always active; views, animations or bond rotations can be built in by the kinemage author. A text window explains the intentions of the author, while a simple hypertext capability allows the reader to jump to the specific view and display objects being described; however, most kinemages can also be successfully understood just by exploring what is available within the graphics window.

Kinemages are suitable for structure browsing or producing static two-dimensional presentation graphics, but those aspects have been kept secondary to effectiveness for interactive visualization and flexibility of author specification. Features and representations have deliberately been chosen to be fast and simple but spatially informative, rather than either showy or traditional, as illustrated by the following examples and their rationales. Mouse-controlled rotation in *KiNG* depends only on the direction of drag, so that the behaviour of the image is independent of absolute cursor position within the window. Labels are available but seldom needed, since the data structure builds in a 'pointID' that is displayed whenever the point is picked. Instead of half-bond colouring, which tends to chop up the image, *Molikin* by default provides separate colours and button controls for main chain *versus* side chains, and it can prepare a partial 'ball & stick' representation with colour-coded balls on non-carbon, non-hydrogen atoms (see Fig. 21.5.1.1). Hydrogen atoms are crucial for some research uses, but to minimize the clutter from twice as many atoms, *Molikin* sets up their display under button control. For effective perception of conformational change, while avoiding either the confusion of overlays or the potential misrepresentation of computed interpolation, *KiNG* features simple animation switching between known conformations. Very importantly, since molecular information resides mostly in chemical bonds and spatial proximity, kinemages emphasize fully three-dimensional representations, such as vectors, dots or balls, rather than surface graphics that obscure internal structure. A space-filling representation (the 'spherelist') is available, but it is suggested that it be used very sparingly – for example, to show the size and shape of a small-molecule ligand. If an extensive surface is needed, a dot surface is more informative, since the underlying atoms and bonds can be seen at the same time. However, a well rendered ribbon is unmatched for conveying overall 'fold'. *Molikin* calculates and *KiNG* displays spline-smoothed ribbon schematics similar to the original hand-drawn style (Richardson, 1981), with black-line edges to aid depth perception (see Fig. 21.5.1.1*a*). For interactive use they serve first as introduction to a structure, and then as context for more detailed ball, vector and dot representations.

Copyright © 2012 International Union of Crystallography

For kinemages, the representation style is not a global choice that applies to everything shown, but rather is a set of local options (varied across space or sequence) chosen to provide appropriate emphasis and comprehensible detail within context.

21.5.2. Uses of *KiNG* and kinemages

21.5.2.1. Use as a reader of existing kinemages

Viewing a pre-existing kinemage file requires almost no learning process: the interface is sufficiently 'transparent' that interaction is mainly with the molecule rather than with the program (Richardson & Richardson, 2002). Six simple operations cover all basic functionalities: (1) drag with the mouse to rotate the displayed object; (2) click on a point to identify it; (3) turn things on or off, or animate if that option is present, with labelled buttons; (4) choose preset views from the Views pull-down menu; (5) read the author's explanations in the text window; and (6) change to the next kinemage in the file with the Kinemage selection list on the upper right corner of the GUI. At a slightly more complex level, one can recentre, zoom the scale, move the clipping planes and save a view; measure distances, angles and dihedrals or 'Find' by point name (from the Edit pull-down menu); change Display menu options such as stereo or perspective; or consult the Help menu. There are keyboard shortcuts for convenience (such as 'a' to animate or 'c' for cross-eye *versus* wall-eye stereo), but they are never the only method and they are defined on the menus.

21.5.2.2. Use for teaching and on the web

Simplicity of interface, free cross-platform availability and attention to presentation issues (such as a keystroke to make labels and measures large and visible for an audience) make *KiNG* and kinemages especially well suited for teaching and learning about macromolecular structure or about crystallographic concepts such as handedness and symmetry. Teachers can use, adapt or create kinemage materials for class- or homework, on the computer or on paper (*e.g.*, file HbAllo.kin on haemoglobin). Interactively evaluated exercises (such as HBondPractice.kin) or timed tests (such as 3Dlit2atst.kin) can be set up in *Mage*. Especially valuable is student construction of reports or projects in kinemage format using *KiNG*, which they find increases their structural understanding and we find lets us evaluate that understanding. Suggestions can be found in Richardson & Richardson (2002) and examples, documentation and links to other kinemage web sites and materials can be found under Kinemages or under Teaching, at the kinemage web site http://kinemage.biochem.duke.edu.

KiNG has a number of options for making illustrations, including direct two-dimensional output of the current graphics window (at enhanced resolution) to JPEG, PNG or PDF image formats, or to POV-Ray for rendering. Movies can be made in *KiNG* of rotation, motion or changes, and either those or the static two-dimensional images can be inserted into PowerPoint-style presentations. Direct use of interactive three-dimensional graphics on the web without need for user installation [as is central to our *MolProbity* (Davis *et al.*, 2007; also Chapter 21.6 of this volume) structure-validation web service] was a major motivation for writing *KiNG* in Java. That same convenience can enhance web materials for courses or the presentation of structural results on a laboratory web site. Examples and simple templates in html are given on the kinemage web site under Software/KiNG.

21.5.2.3. Use for research

For general molecular-structure studies, kinemages act as a three-dimensional laboratory notebook where author and reader are the same person. These kinemages keep a visual record of the

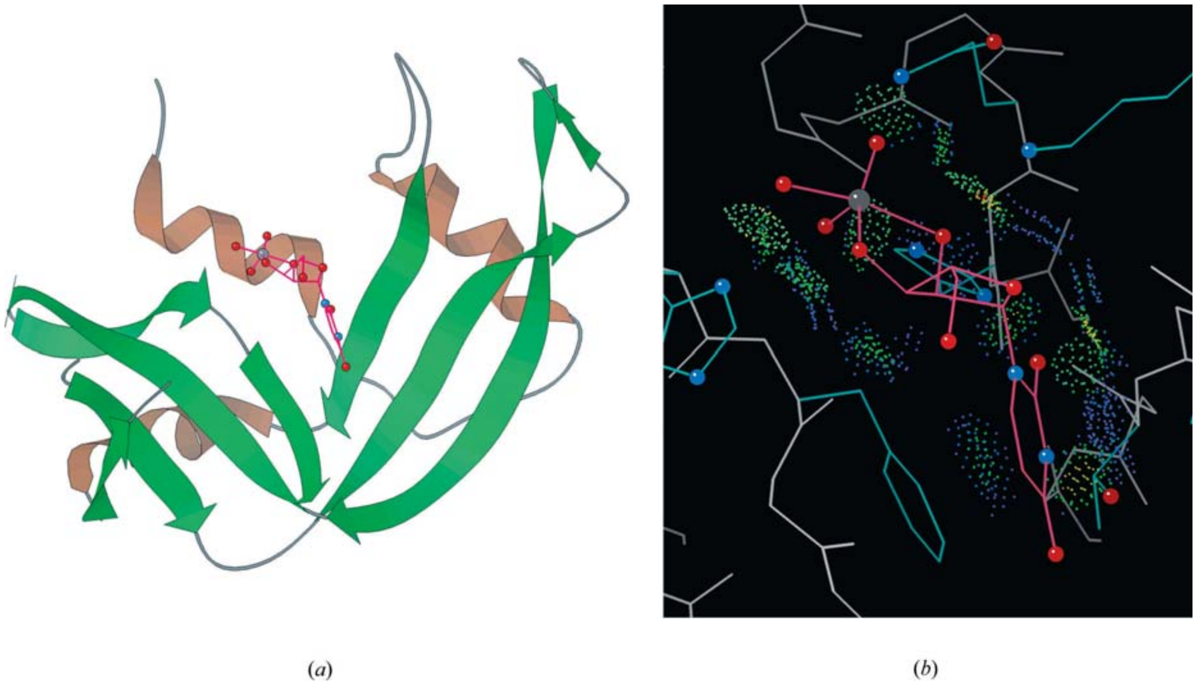
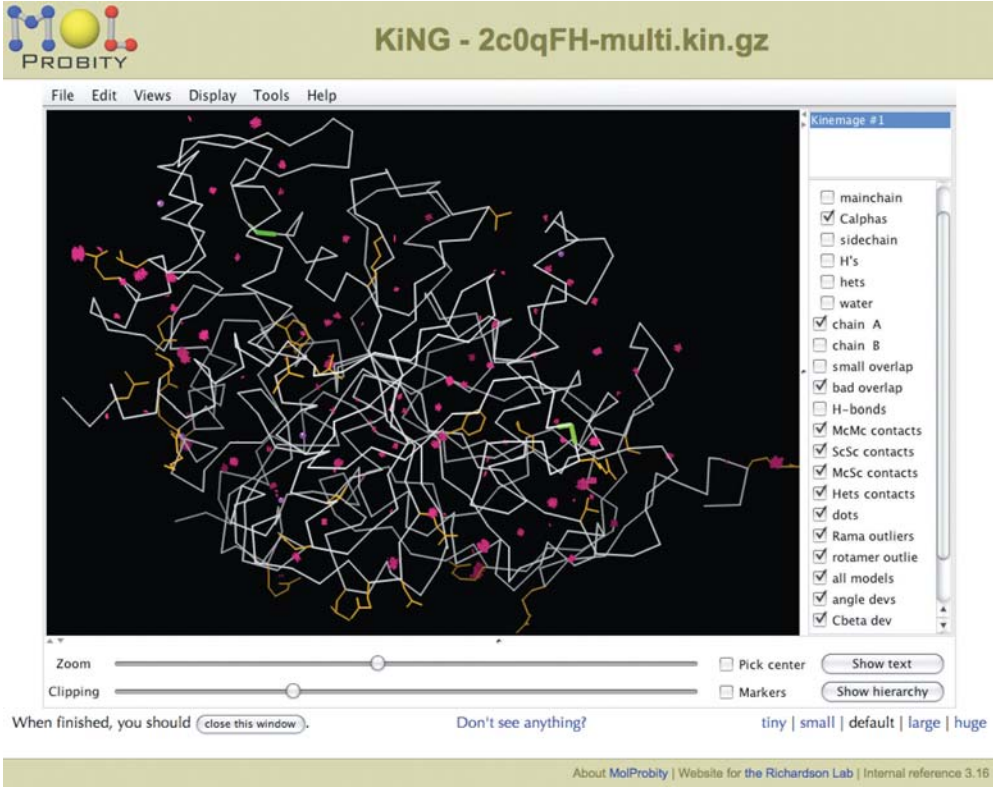


Figure 21.5.1.1
Illustration of combining molecular representations in kinemage format, shown in *KiNG*. (*a*) shows a ribbon schematic of the uridine vanadate–ribonuclease A complex at 1.3 Å resolution (PDB code: 1RUV) (Ladner *et al.*, 1997), with the bound inhibitor in ball representation. (*b*) shows a close up of the binding pocket, with the contacting side chains (cyan) and backbone (white) of the enzyme as vectors, the uridine vanadate as balls, and all-atom contacts at the enzyme–inhibitor interface as drawn by *Probe*; hydrogen bonds are pillows of pale green dots, favourable van der Waals contacts are outlined by blue, green and yellow dots, and there are no steric clashes.



PROBITY
KiNG - 2c0qFH-multi.kin.gz
File Edit Views Display Tools Help
Kinemage #1
Zoom
Clipping
Show text
Show hierarchy
When finished, you should close this window.
Don't see anything?
tiny | small | default | large | huge
About MolProbity | Website for the Richardson Lab | Internal reference 3.16

☐ mainchain
☑ Calphas
☐ sidechain
☐ H's
☐ hets
☐ water
☑ chain A
☐ chain B
☐ small overlap
☑ bad overlap
☐ H-bonds
☑ McMc contacts
☑ ScSc contacts
☑ McSc contacts
☑ Hets contacts
☑ dots
☑ Rama outliers
☑ rotamer outlie
☑ all models
☑ angle devs
☑ Cbeta dev

☐ Pick center
☐ Markers

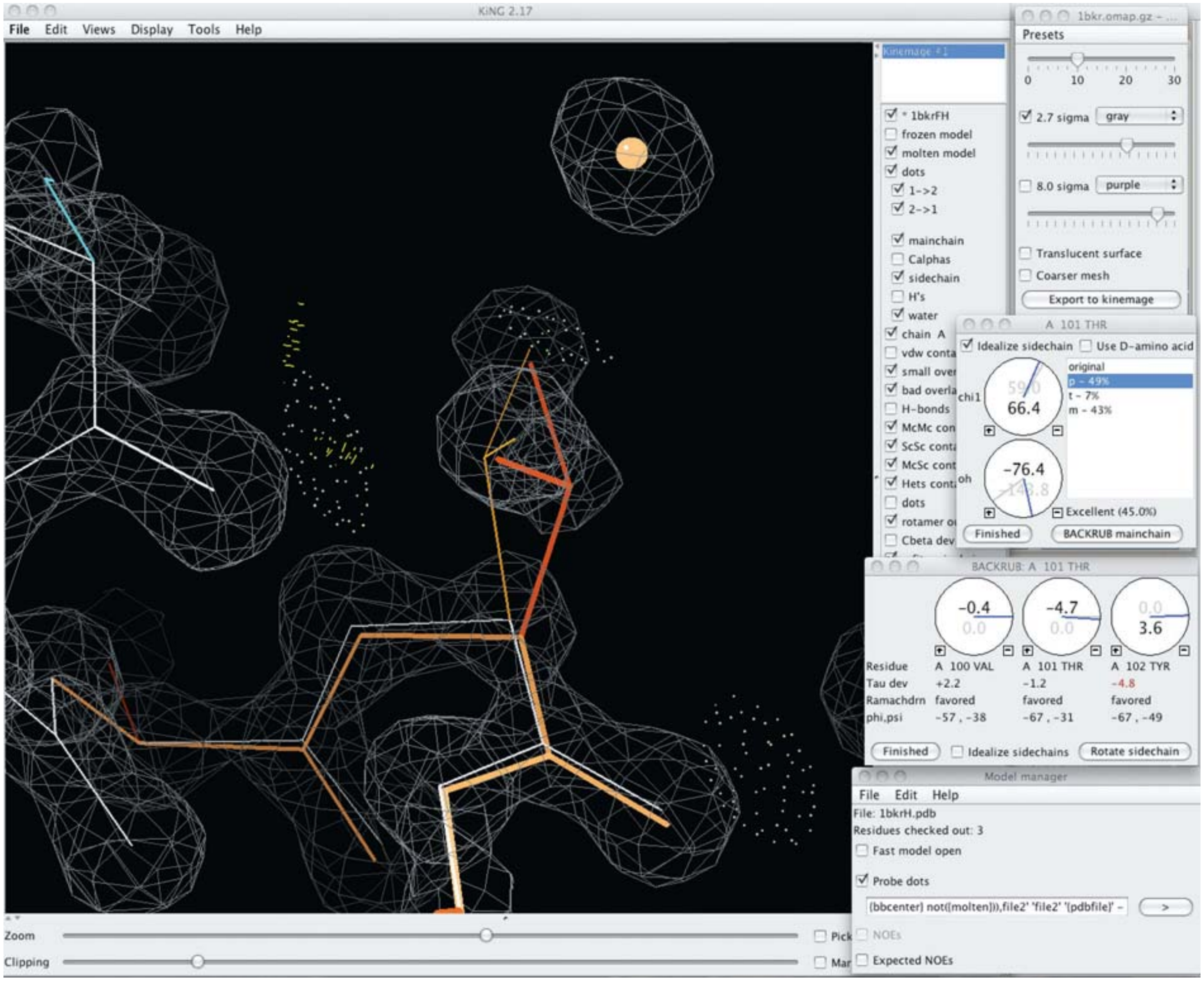
Figure 21.5.2.1
Screenshot of *KiNG* presenting interactive three-dimensional validation data directly in the browser, in a *MolProbity* session for PDB code 2C0Q (Ekstrom *et al.*, 2006), a better-than-average 2.5 Å resolution structure. On the Cα backbone (white) all validation outliers are shown: clashes as clusters of hot-pink spikes, Ramachandran outliers in green, rotamer outliers as gold side chains and Cβ deviations as magenta spheres; no bond-length or angle outliers are visible. The user can centre, zoom in and turn on all atoms to study a given problem in detail.

research process with selections, views, labels, measurements, superpositions *etc.*, plus a descriptive record in the text window. On-screen editing tools allow easy deletion of unwanted parts, constructions such as ring centres or helix axes and measurements on them, drawing shortened or dotted lines, recolouring or dragging points or regions, and many other operations in addition to the molecular rebuilding described below.

Setting up an animation between conformations or between related structures is an easy and very sensitive way of seeing changes, including correlated motions. Completely new display objects and organizations can be added to kinemages, such as three-dimensional plots of related non-molecular data. Kinemages are an easy and platform-independent way of sharing ideas with collaborators, either side-by-side or at a distance with simultaneous discussion, or just by sending a kinemage with its preset views and notes. [This is easy enough that a co-laboratory function through socket connection in *Mage* has rarely been used, and was therefore not implemented in *KiNG*.] Later, the working research kinemages can be used to produce either static two-dimensional or interactive illustrations for lectures or publication.

KiNG has been useful for displaying data other than molecular structure, however, since the kinemage format is a generic plotting format, not tied to display of molecules. For instance, the *MolProbity* structure validation web server (Davis *et al.*, 2007; Chen *et al.*, 2010; Chapter 21.6) generates two-dimensional Ramachandran data plots and three-dimensional Cβ deviation scatter plots which can be downloaded and viewed in *KiNG*. In addition, support for higher-dimensional data has been added to *Mage* and *KiNG*. For high-dimensional kinemages, the user chooses on the fly which three dimensions to display at a time. Also, a parallel-coordinate view (Inselberg, 1985) is available with a single keystroke, which displays the multiple axes as a series of parallel lines and each multi-dimensional point as a poly-line across those axes. With this combination of features, users can track, colour and measure the behaviour of clusters of points in many dimensions. Seven-dimensional kinemage analysis was particularly instrumental in defining full-detail RNA backbone conformers (Richardson *et al.*, 2008).

Two important tools for assessing the quality of structures are *Probe* and *Reduce*. *Probe* analyses molecular interactions by calculating small-probe contact dots wherever two atoms are within 0.5 Å of van der Waals contact (Word, Lovell, LaBean *et al.*, 1999), for numerical scoring or for display in *KiNG* (Fig. 21.5.1.1*b*), where the three types of contacts (hydrogen bonds, favourable van der Waals contacts and unfavourable 'clash' overlaps) are under separate control. All-atom contact analysis requires all hydrogen atoms; they are added by *Reduce* (Word, Lovell, Richardson & Richardson, 1999; Davis *et al.*, 2007), which optimizes the orientations of OH, SH and NH_3 groups, and possible 180° flips of Asn, Gln or His side-chain ends, considering both van der Waals clashes and hydrogen bonds analysed combinatorially in local networks. These contact-surface tools have research uses that fall into two distinct categories: one is study of the patterns and causes of particular structural features in molecules (best done on atomic resolution structures; Fig. 21.5.1.1*b*); the other is sensitive testing, validating and adjusting of an individual molecular model, either computational or experimental (exemplified by the *MolProbity* validation analysis).



☑ * 1bkrFH
☐ frozen model
☑ molten model
☑ dots
☑ 1->2
☑ 2->1
☑ mainchain
☐ Calphas
☑ sidechain
☐ H's
☑ water
☑ chain A
☐ vdw conta
☑ small over
☑ bad overla
☐ H-bonds
☑ McMc con
☑ ScSc conta
☑ McSc cont
☑ Hets conta
☐ dots
☑ rotamer o
☐ Cbeta dev

☑ 2.7 sigma gray
☐ 8.0 sigma purple
☐ Translucent surface
☐ Coarser mesh

☑ Idealize sidechain ☐ Use D-amino acid
chi1 59.0 66.4
oh -76.4 -143.8
original
p - 49%
t - 7%
m - 43%
Excellent (45.0%)

	-0.4 0.0	-4.7 0.0	0.0 3.6
Residue	A 100 VAL	A 101 THR	A 102 TYR
Tau dev	+2.2	-1.2	-4.8
Ramachdrn	favored	favored	favored
phi,psi	-57 , -38	-67 , -31	-67 , -49

☐ Idealize sidechains

File: 1bkrH.pdb
Residues checked out: 3
☐ Fast model open
☑ Probe dots
{bbcenter} not({molten})),file2' 'file2' '{pdbfile}' –
☐ NOEs
☐ Expected NOEs
☐ Pick
☐ Mar

Figure 21.5.2.2
Screen capture of side-chain rebuilding in *KiNG* using the side-chain rotator and backrub tools, with display of electron density, all-atom contacts and rotamer quality. In this example from the 1.1 Å resolution crystal structure of the calponin homology domain of β-spectrin (PDB code: 1BKR) (Banuelos *et al.*, 1998), a threonine was fitted with its side chain flipped over by 180°. The original (in gold since it is a rotamer outlier) has large bond-angle outliers and steric clashes, and its Cβ is entirely out of density. Using the side-chain rotator tool in combination with the backrub tool allows a much better fit to the density (the thicker orange 'mobile' side chain), with an excellent rotamer, ideal geometry and two hydrogen bonds (pale green dot pillows) for the Thr OH. H atoms and van der Waals dots are turned off, for clarity in two dimensions.

21.5.2.4. *Cystallographic rebuilding tools*

In crystallography, the most important use of all-atom contact analysis is for quickly finding, and frequently for fixing, problems with molecular geometry during fitting and refinement. All-atom contacts add independent new information to that process, since the hydrogen atoms make almost no contributions either to the electron density or to the energetic component of refinement as presently done, yet they are undeniably present and cannot significantly overlap other atoms (except in hydrogen bonds). The steric constraints implied by all-atom contacts are significantly more stringent than those based only on non-hydrogen atoms, yet they are obeyed almost perfectly by low-*B* regions of structures at resolutions near 1 Å, even when hydrogen atoms were not used in refinement.

At any stage of a structure determination, contact dots for the entire molecule or molecules can be calculated by *Probe* and examined in *KiNG*, or a list of the severe clashes and other outliers can be generated, for printout or as a scripted 'to-do' list for rebuilding in *Coot* (Emsley & Cowtan, 2004; Emsley *et al.*, 2010).

The most common way to do *MolProbity* analysis on a structure is on the web site (http://molprobity.biochem.duke.edu) (Davis *et al.*, 2007; Chen *et al.*, 2010; Chapter 21.6), which guides the entire process through to both chart and kinemage analysis of the local quality of a macromolecular structure. The resulting 'multi-criterion' kinemage flags all the individual outliers on the three-dimensional structure (Fig. 21.5.2.1): serious clashes ($\geq$0.4 Å overlap), poor side-chain rotamers, backbone bond-length or angle outliers, Cβ deviations, Ramachandran outliers and, for RNA, sugar-pucker and backbone conformer outliers (Richardson *et al.*, 2008). That kinemage can be viewed in *KiNG* online, but for rebuilding to correct the diagnosed problems a user would download the multi-criterion kinemage and the H- and flip-optimized PDB file for off-line work in *KiNG*. The multi-criterion kinemage shows where work is needed. However, the most powerful use of all-atom contact information is directly and interactively in the process of fitting and rebuilding. Therefore, the rebuilding tools in *KiNG* can call *Probe* on-the-fly during refitting, for display of clashes, hydrogen bonds and favourable van der Waals contacts (as well as updated rotamer quality and checks for φ, ψ and τ outliers) (Lovell *et al.*, 2000, 2003). This gives instant feedback on how user-proposed changes improve or worsen the model locally. These tools, including a side-chain mutation tool, a side-chain rotator tool, a backrub tool and an RNA rotator tool for adjusting RNA suite conformations, read in PDB files (with H atoms) and save all user modifications directly to PDB format files, for ease of use with other refinement or analysis programs. *KiNG* can also read in and display most common formats of electron density or difference density files, which are essential to the rebuilding process. Side chains are by default added in ideal geometry and in a choice of ideal rotamers, with user adjustment of torsion angles from there. The 'backrub' motion (Davis *et al.*, 2006) is a subtle rotation of the dipeptide, shown to occur frequently in protein molecules and especially valuable in side-chain modelling because of its leverage on the Cα–Cβ orientation. Fig. 21.5.2.2 shows an example of the side-chain rotator, backrub and real-time contact tools in action.

A number of other commonly used crystallographic tools have been modified to use *Probe* and *Reduce* as well. Historically, both *O* (Jones *et al.*, 1991) and *XtalView* (McRee, 1999) included some facility for all-atom *Probe* analysis. *Coot* (Emsley & Cowtan, 2004; Emsley *et al.*, 2010) can call *Reduce* and *Probe* directly, displaying hydrogens and all-atom contact dots during refitting. Also, *MolProbity* can write a scheme or Python input script for *Coot* which produces a list of labelled buttons that take the *Coot* user to each residue with a *MolProbity*-diagnosed clash, Ramachandran or rotamer outlier. Most notably, all *MolProbity* evaluations are available in *Phenix* (Adams *et al.*, 2010), with display in *KiNG* and interactive links for outlier corrections in *Coot*.

In *KiNG* or in these other systems, using all-atom contact display during refitting often makes conformational choices unambiguous, even when the electron density alone does not distinguish them. This criterion can locate a Met methyl, find the correct orientation for the final branch of an Asn or Gln, or a Thr, Val or Leu, improve the backbone conformation of a Gly, disentangle alternate conformations, or show which direction a ligand binds, all at a lower resolution than otherwise possible.

Side-chain rebuilding in *KiNG* is also very effective for analysing the suitability and probable effect of single-site mutations or redesigns (*e.g.* Ghaemmaghami *et al.*, 1998; Grell *et al.*, 2000; Wales *et al.*, 2004; Humphris & Kortemme, 2008). That effectiveness depends on the interactive update of rotamer quality and all-atom contacts, and is enhanced by use of the backrub motion, originally in *KiNG* and now also in *Coot* and several protein design systems (Georgiev *et al.*, 2008; Smith & Kortemme, 2008).

21.5.3. Making kinemages

There are numerous methods of creating kinemages. At the most basic level, users can hand-type a set of three-dimensional coordinates into a text file, adding only a small amount of hierarchical organization, and then view the data. Both *KiNG* and *Mage* by policy are very tolerant to parsing alternatives, minor mistakes and missing objects within a kinemage; so as long as an input file is reasonably close to being correct, something will still be displayed. Naturally, though, for large amounts of data, hand-creation of kinemages is impractical, so for viewing three-dimensional (or higher-dimensional) statistical data, *KiNG* includes a 'Data plotter' tool which will read in a delimited spreadsheet text file and automatically create a high-dimensional kinemage from those data.

For creating kinemages from PDB or mmCIF files, *KiNG* has several options. *Molikin* allows *KiNG* to read PDB or mmCIF files directly from the command line, with a small number of command-line flags for specifying some of the most common kinemage options, including ribbons or 'lots' (which produces vectors for the main chain, side chains, hydrogen atoms and non-water heterogens, small balls for waters, and pointIDs that include *B* factors). In the *KiNG* GUI, the 'Import' tool in the File menu provides access to *Molikin* features as well. For quick browsing of structure files, drag-n-drop or the 'Import' → 'To quick kin' menu allows users to quickly import and construct kinemages, similar to the command-line options. For more direct control of the options shown in the kinemage, the 'Import'→ 'Molecules' option brings up the *Molikin* GUI, which allows users to choose precisely what is shown. Users can specify what models, chains, residues, side-chain types and atoms to show, in a variety of representation types.

The many modes and options make *KiNG* and kinemages useful for a myriad of purposes, in addition to simple structure browsing. Making a kinemage for research, teaching, publication or distribution is an iterative and deliberate process. While the built-in kinemages provide a good starting point, finding the optimal colours, views and objects to display can be time consuming. *KiNG* includes a number of built-in on-screen editing tools which make this task easier.

On-screen editing of a kinemage in *KiNG* usually begins with setting up a few good views: rotate, pick centre, zoom and clip to optimize each one, and save it with 'Save current view' on the Views pull-down menu; it then shows up on the same menu with the given label. For editing of object properties in the kinemage, the 'Edit/draw/delete' tool in the Tools menu brings up a separate toolbox containing a variety of editing tools. These options, once selected, utilize mouse clicks in the graphics window to determine the objects that are modified, often with additional options that only appear when the particular option is selected. For example, the 'Paint points' option brings up a colour selection drop-down list, as well as a selection-type list. By default, the 'Paint points' option allows users to paint all the points that appear within a circle, which is drawn in the GUI window to allow users to see precisely what they are painting before they actually paint. A variety of drawing and deletion tools are also included in the 'Edit/draw/delete' tool. 'Edit text' (Edit menu) brings up the text window, which allows explanations to be included with the kinemage. Several superposition or docking tools, at varying levels of complexity, allow the user to move one molecule or group onto another similar one, for comparison or animation.

KiNG includes a built-in 'Internal palette' kinemage (in the Help menu) which shows the kinemage palette of colours with their names and gives some guidelines for choosing effective colours. The palette includes emphasis colours (green, yellow, hot pink), low-emphasis colours (purple, brown), pastels and a large variety of mid-tones (such as gold, peach, lilac), with a depth-cueing formula that ensures smooth change and maintenance of the perceptual colour value at all depths. In *KiNG*, if the built-in palette does not suffice, the user can create custom HSV (hue, saturation, value) colours. The default black background works best on the computer screen or for lecture-hall projection, but a white-background mode with a modified palette is available when needed for printed images (*e.g.* Fig. 21.5.1.1*a*) or for projection in high-ambient-light conditions.

Context is important for a kinemage (usually at least overall C$^{\alpha}$'s) but often it is helpful to delete features that are not directly relevant to the current interest (*e.g.* Fig. 21.5.1.1); therefore *KiNG* includes a versatile set of deletion tools (punch, prune, auger, spherical crop). This selection process is like the simplification and emphasis needed for a good two-dimensional illustration, but in this case it applies to the fully interactive three-dimensional form. For a kinemage, however, it is both possible and advantageous to include some additional details for further exploration, controlled by a button which can start out turned off.

At any stage of kinemage editing, a text editor can be used to look at the plain ASCII kinemage file, with its text, its views, and the hierarchy of group, subgroup and list display objects in human-readable and clearly identified forms. Lists (*e.g.* @vectorlist {name}) can be of vectors, dots, labels, words, balls, spheres, triangles or ribbons. Any part of the file can be edited, using its existing format as a guide or looking at another kinemage file that provides a desired template. The formal rules are described in format-kinemage.pdf, linked from Software/KiNG on the kinemage web site. One of the few operations that currently must be edited outside rather than inside *KiNG* is moving points between different lists (for instance, setting up a new list of just active-site side chains in a different colour and controlled by their own button). After editing a file outside *KiNG*, the kinemage should be saved without formatting, as a plain text file.

All in all, making a simple kinemage is trivial, but making really good ones for use by others is much like making a good web page. There are tools that make the individual steps easy, but one needs to exercise restraint to keep it simple enough to be both fast and comprehensible, patience to keep looking at the result and modifying it where needed, and judgment about both content and aesthetics.

21.5.4. Software notes

Mage and *Prekin* were written in C by David C. Richardson. *KiNG* and *Molikin* were written in Java by Ian W. Davis and Vincent B. Chen with contributions by Daniel A. Keedy, and are compatible with any Java 1.5+ equipped system. *Probe* (in C) and *Reduce* (in C++) were written by J. Michael Word and updated by Andrew Leaver-Fay and other collaborators in Jack Snoeyink's group. The latest versions of all the software, compiled for multiple platforms, plus source and documentation files are available free from the Software section of the kinemage web site (http://kinemage.biochem.duke.edu). The *KiNG* documentation is included in its software download package, and is also available there separately as file king-manual.pdf.

Kinemage development is supported by NIH grants GM073919 and GM 73930, and historically by the Protein Society and by GM15000.

References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. *Acta Cryst.* D**66**, 213–221.

Banuelos, S., Saraste, M. & Carugo, K. D. (1998). *Structural comparisons of calponin homology domains: implications for actin binding*. *Structure*, **6**, 1419–1431.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Chen, V. B., Arendall, W. B. III, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *MolProbity: all-atom structure validation for macromolecular crystallography*. *Acta Cryst.* D**66**, 12–21.

Chen, V. B., Davis, I. W. & Richardson, D. C. (2009). *KiNG (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program*. *Protein Sci.* **18**, 2403–2409.

Davis, I. W., Arendall W. B. III, Richardson J. S. & Richardson D. C. (2006). *The backrub motion: how protein backbone shrugs when a sidechain dances*. *Structure*, **14**, 265–274.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids*. *Nucleic Acids Res.* **35**, W375–W383.

Ekstrom, F., Akfur, C., Tunemalm, A. K. & Lundberg, S. (2006). *Structural changes of phenylalanine 338 and histidine 447 revealed by the crystal structures of tabun-inhibited murine acetylcholinesterase*. *Biochemistry*, **45**, 74–81.

Emsley, P. & Cowtan, K. (2004). *Coot: model-building tools for molecular graphics*. *Acta Cryst.* D**60**, 2126–2132.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Features and development of Coot*. *Acta Cryst.* D**66**, 486–501.

Georgiev, I., Keedy, D., Richardson, J. S., Richardson, D. C. & Donald, B. R. (2008). *Algorithm for backrub motions in protein design*. *Bioinformatics*, **24**, i196–i204.

Ghaemmaghami, S., Word, J. M., Burton, R. E., Richardson, J. S. & Oas, T. G. (1998). *Folding kinetics of a fluorescent variant of monomeric l repressor*. *Biochemistry*, **37**, 9179–9185.

Grell, D., Richardson, J. S., Richardson, D. C & Mutter, M. (2000). *SymROP: a ROP protein redesign with four identical helices, for TASP synthesis*. *J. Mol. Graphics Model.* **18**, 290–298.

Humphris, E. L. & Kortemme, T. (2008). *Prediction of protein–protein interface sequence diversity using flexible backbone computational protein design*. *Structure*, **16**, 1777–1788.

Inselberg, A. (1985). *The plane with parallel coordinates*. *Vis. Comput.* **1**, 69–91.

Jones, A. T., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Improved methods for building protein models in electron density maps and the location of errors in these models*. *Acta Cryst.* A**47**, 110–119.

Ladner, J. E., Wladkowski, B. D., Svensson, L. A., Sjölin, L. & Gilliland, G. L. (1997). *X-ray structure of a ribonuclease A–uridine vanadate complex at 1.3 Å resolution*. *Acta Cryst.* D**53**, 290–301.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Structure validation by* $C\alpha$ *geometry:* φ, ψ *and* $C\beta$ *deviation*. *Proteins*, **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000). *The penultimate rotamer library*. *Proteins*, **40**, 389–408.

McRee, D. E. (1999). *XtalView/Xfit – a versatile program for manipulating atomic coordinates and electron density*. *J. Struct. Biol.* **125**, 156–165.

Richardson, D. C. & Richardson, J. S. (1992). *The kinemage: a tool for scientific illustration*. *Protein Sci.* **1**, 3–9.

Richardson, D. C. & Richardson, J. S. (1994). *Kinemages – simple macromolecular graphics for teaching and publication*. *Trends Biochem. Sci.* **19**, 135–138.

Richardson, D. C. & Richardson, J. S. (2002). *Teaching molecular 3-D literacy*. *Biochem. Molec. Biol. Educ.* **30**, 21–26.

Richardson, J. S. (1981). *The anatomy and taxonomy of protein structures*. *Adv. Protein Chem.* **34**, 167–339.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J. & Berman, H. M. (2008). *RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)*. *RNA*, **14**, 465–481.

Smith, C. A. & Kortemme, T. (2008). *Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction*. *J. Mol. Biol.* **380**, 742-–756.

Wales, T. E., Richardson, J. S. & Fitzgerald, M. C. (2004). *Facile chemical synthesis and equilibrium unfolding properties of CopG*. *Protein Sci.* **13**, 1918–1926.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms*. *J. Mol. Biol.* **285**, 1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation*. *J. Mol. Biol.* **285**, 1735–1747.

International Tables for Crystallography (2012). Vol. F, Chapter 21.6, pp. 694–701.

Chapter 21.6. *MolProbity*: all-atom structure validation for macromolecular crystallography

V. B. Chen, W. B. Arendall III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson

21.6.1. Summary of *MolProbity* flow and user interactions

The usual interaction with *MolProbity* (Davis *et al.*, 2007) is through the internet at http://molprobity.biochem.duke.edu or as a main menu item on our general laboratory website at http://kinemage.biochem.duke.edu. [For bulk users, it is also possible to set up your own local *MolProbity* server or to use the individual programs in command-line mode. *MolProbity*-style validation is also now available in *PHENIX* (see Section 21.6.6).] Tutorial exercises for the whole process of diagnosing and fixing errors can be found on the kinemage site under Teaching/MolProbity.

A typical *MolProbity* session starts with the user uploading a coordinate file of their own or fetching one from the PDB or NDB databases (Berman *et al.*, 1992, 2000) in new or old PDB format or in mmCIF format. After checking the thumbnail image and listed characteristics of the input file, and editing or reloading if needed, H atoms are added and optimized, with automated correction of Asn/Gln/His 180° flips if needed (Word, Lovell, Richardson & Richardson, 1999).

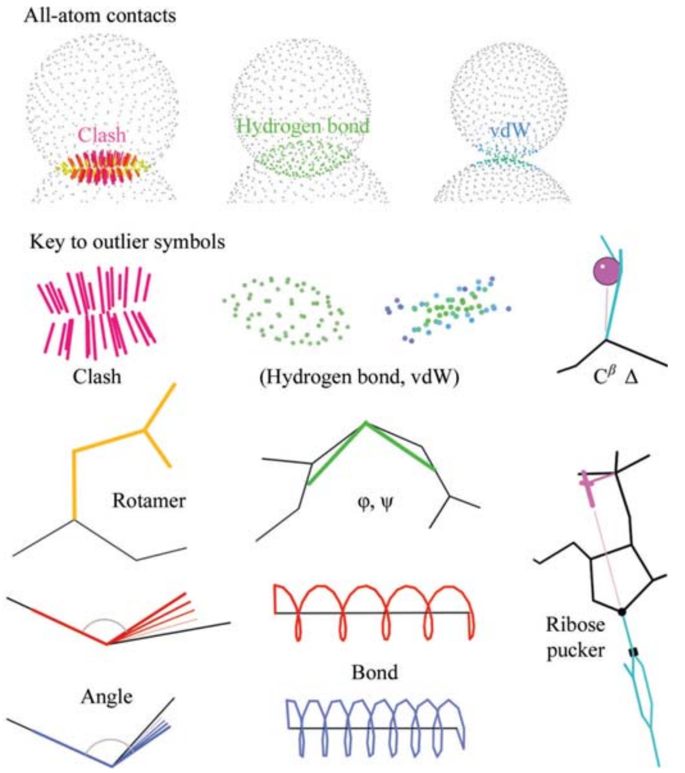


Figure 21.6.1.1
An outlier legend, showing each symbol used in a *MolProbity* multi-criterion kinemage and illustrating the relationship of the three types of all-atom contact to the atomic van der Waals (vdW) surfaces (spheres of small grey dots). The symbols for favourable hydrogen bonds and vdW contacts are included for completeness, as well as the hot-pink spikes of a clash outlier. A C^{β} deviation of ≥ 0.25 Å is shown as a magenta ball centred on the ideal C^{β} position and tangent to the modelled position. Bad rotamers are shown as gold side chains and Ramachandran outliers as heavy green lines to the midpoints of the two peptides. Bond-angle outliers are indicated by a fan of lines from the ideal to the modelled bond (red if wide, blue if narrow). Bond-length outliers are indicated as stretched (red) or compressed (blue) springs. A suspicious ribose pucker is diagnosed by the perpendicular distance from the 3′ (following) phosphate to the line of the glycosidic C1′—N1/9 bond and is flagged by a representation of that construction (in magenta if too short, as here, and in purple if too long).

The user then chooses which validation analyses to run and what reports and output files to generate. The *MolProbity* interface adjusts the defaults and options presented and even the page flow depending on user choices and on the properties of the file being worked on. These adjustments make *MolProbity* simple for novice users, while at the same time allowing advanced users to have great control over their runs. The core 'glue' that generates the HMTL code controlling the main user interface and programmatic interactions of *MolProbity* is implemented in the PHP programming language. Underlying the PHP core, the majority of the analysis tasks in *MolProbity* are performed by individual programs written in a range of languages, including C, C++, Java and Perl. It uses *REDUCE* and *PROBE* for all-atom contact analysis, *RAMALYZE*, *ROTALYZE*, *DANGLE*, *SILK* and *SUITENAME* for other criteria and *KiNG* (see Chapter 21.5) for three-dimensional visualization of the structure and its validation markers directly in the browser. Fig. 21.6.1.1 shows a key to *MolProbity*'s graphical markers for validation outliers. Further details are provided below on the specific analyses that *MolProbity* can perform. The validation results are reported in the form of summaries, charts, two-dimensional and three-dimensional graphics and output files for download.

The crucial final step in the *MolProbity* process is for the crystallographer to download the result files and work off-line to correct as many of the diagnosed problems as feasible. Rebuilding with consideration of the validation outliers, the electron density and the surrounding model is usually performed either in *Coot* (Emsley *et al.*, 2010) or in *KiNG* (Chen *et al.*, 2009; Chapter 21.5). At resolutions of about 2.5 Å or better it is possible to correct the great majority of outliers (Arendall *et al.*, 2005), with an order-of-magnitude improvement in the various *MolProbity* scores and some improvement in geometry, map quality, R factor and $R_{\rm free}$. An example is shown in Fig. 21.6.1.2 with before-and-after multi-criterion kinemages.

21.6.2. Validation analyses

21.6.2.1. *Addition of H atoms*

The presence of H atoms (both nonpolar and polar) is a critical prerequisite for all-atom contact analysis. Although refinement using H atoms is becoming more common, most crystal structures are still deposited without H atoms. Once a coordinate file has been uploaded, *MolProbity* detects whether the file contains a suitable number of H atoms; if not, then the 'Add H atoms' option is presented to users first. *MolProbity* uses the software *REDUCE* (Word, Lovell, Richardson & Richardson, 1999; Davis *et al.*, 2007) to add and optimize hydrogen positions in both protein and nucleic acid structures, including ligands, but does not add explicit H atoms to waters. OH, SH and NH_3 groups (but

Copyright © 2012 International Union of Crystallography

not methyl groups) are rotationally optimized and His protonation is chosen within each local hydrogen-bond network, including interactions with the first shell of explicit waters.

A common problem is that the side-chain ends of Asn, Gln and His are easily fitted 180° backwards, since the electron density alone cannot usually distinguish the correct choice of orientation. *REDUCE* can automatically diagnose and correct these types of systematic errors by considering all-atom steric overlaps as well as hydrogen bonding within each local network. Automatic correction of Asn/Gln/His flips is the default option in *MolProbity* during addition of H atoms. *MolProbity* presents each potential flip correction to the user in kinemage view so they have the option of inspecting the before-and-after effects of each flip and approving (or rejecting) each correction. Fig. 21.6.2.1 shows an example of a simple Gln flip that is unquestionably correct but that could not have been decided on the basis of hydrogen bonding alone. Other examples can be much more complex, with rotatable OH positions, large hydrogen-bond networks and multiple competing interactions evaluated exhaustively.

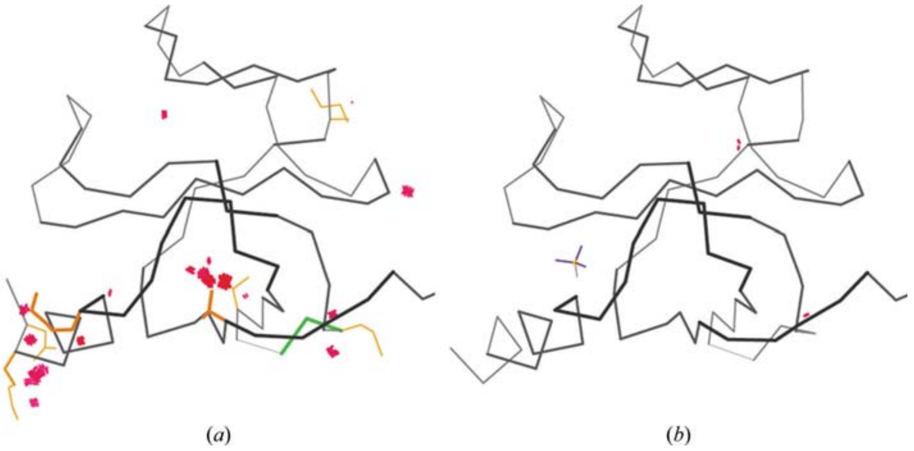


Figure 21.6.1.2
Two multi-criterion validation kinemages illustrating the successful outcome of an overall process of *MolProbity* diagnosis and structure improvement. (*a*) The original 1lpl Cap-Gly structure (Li *et al.*, 2002) shows three major clusters of clash, rotamer and Ramachandran problems plus a few isolated outliers. (*b*) The corrected 1tov structure (Arendall *et al.*, 2005) has essentially no outliers, a 4% lower $R_{\rm free}$, a bound sulfate and an additional turn of helix at the N-terminus.

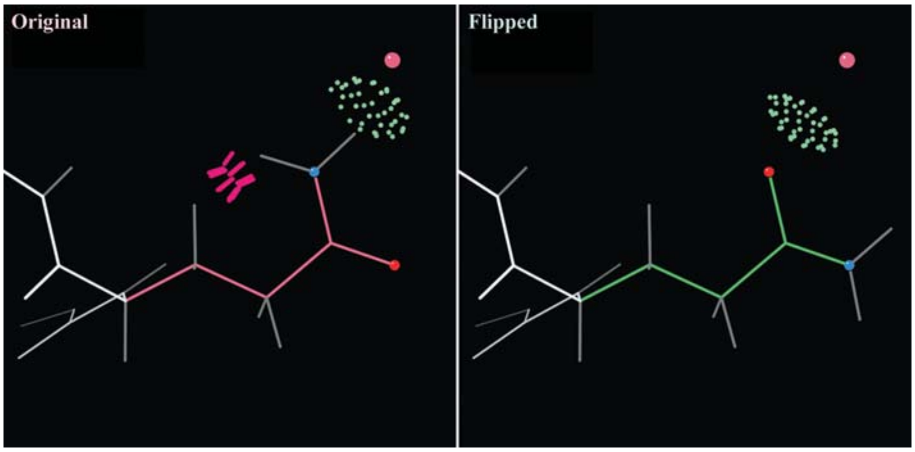


Figure 21.6.2.1
The simple 'flip' correction of a Gln side-chain amide in the 2dq4 threonine 3-dehydrogenase structure (R. Omi, T. Yao, M. Goto, I. Miyahara & K. Hirotsu, unpublished work), a better-than-average 2.5 Å resolution structure. Both orientations make a hydrogen bond to the crystallographic water, but the original has a serious internal clash of the NH_2 group with its own C^{β} hydrogen.

Users can also choose to add H atoms without Asn/Gln/His flips, which is useful in evaluating the atomic coordinates as they were deposited, but which rejects the easiest and most robustly correct improvement that can be made in a crystallographic model (Word, Lovell, Richardson & Richardson, 1999; Higman *et al.*, 2004). If flips are performed, the user needs to download and use the corrected PDB file (either with or without the H atoms) in order to benefit.

21.6.2.2. All-atom contact analysis

Once H atoms have been added to (or detected in) a structure, then the complete 'Analyze all-atom contacts and geometry' option is enabled. A main feature of this option is the all-atom contact analysis, which is performed by the program *PROBE* (Word, Lovell, LaBean *et al.*, 1999). *PROBE* operates by, in effect, rolling a 0.5 Å diameter ball around the van der Waals surfaces of atoms to measure the amount of overlap between pairs of nonbonded atoms. When non-donor–acceptor atoms overlap by more than 0.4 Å, *PROBE* denotes the contact as a serious clash, which is included in the reported clashscore and is shown in kinemage format as a cluster of hot-pink spikes in the overlap region (Fig. 21.6.1.1). Such large overlaps cannot occur in the actual molecule, but mean that at least one of the two atoms is modelled incorrectly. *MolProbity* allows users to select any combination of clashes, hydrogen bonds and van der Waals contacts to calculate and display on the structure. By default, all three are enabled for structures that are not excessively large; for large structures, van der Waals contacts are deselected.

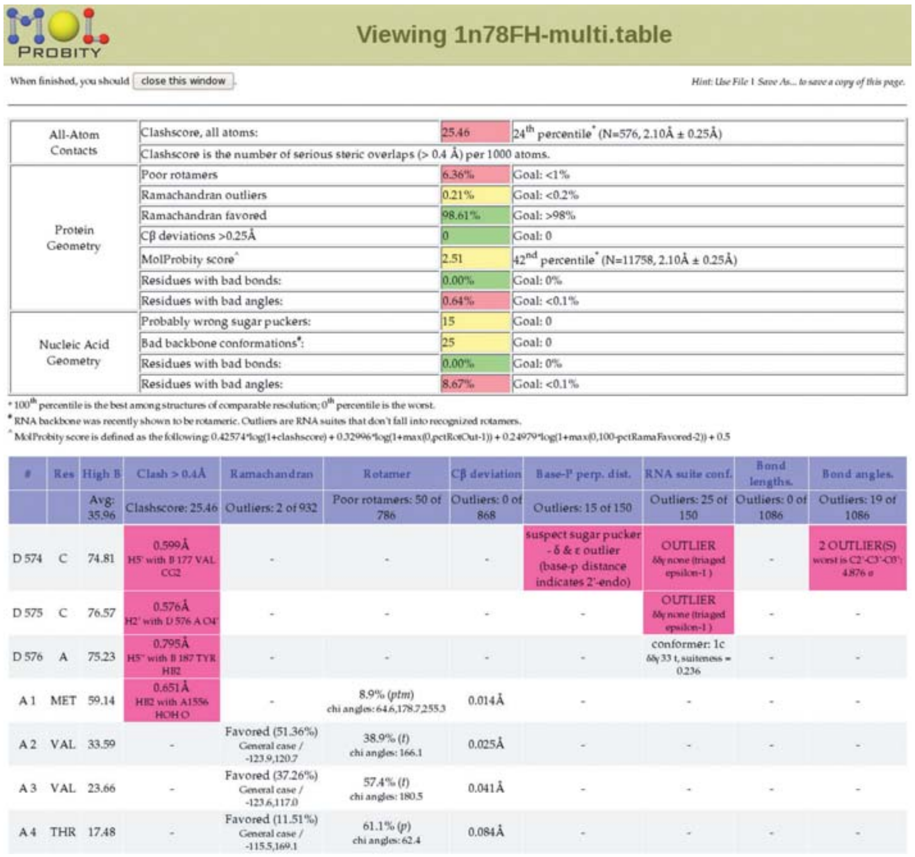
The 'clashscore' is the number of serious clashes per 1000 atoms. It is reported in the *MolProbity* summary (top of Fig. 21.6.2.2), with a red/yellow/green colour coding for absolute quality. The structure's percentile rank for clashscore value within the relevant resolution range is also given. In the detailed sortable 'multi-chart' (an extract is shown below the summary in Fig. 21.6.2.2), the worst clash ≥0.4 Å is listed for each residue and highlighted in pink.

21.6.2.3. Torsion-angle combinations: updated Ramachandran and rotamer analyses

Also included in the 'Analyze all-atom contacts and geometry' option is an evaluation of where residues fall in the multi-dimensional distributions of Ramachandran backbone φ, ψ angles and side-chain rotamer χ angles. The reference distributions are currently from 100 000 residues in 500 files, quality-filtered at both the file and the residue level. The Ramachandran plots are separated for Gly, Pro and pre-Pro residue types; the general plot has only one in 2000 residues outside the 'allowed' contour, which is the same probability as a 3.5σ outlier in a normal distribution. The three specific plots can be robustly contoured only down to excluding one in 500 residues (about 3σ) in the current reference data, but will soon be updated. By ‘robust’ we mean that the contour does not shift with further improvement in resolution or B or with different subselections of the data. When values plateau in this way we can define clear absolute goals for the measure, such as 98% for Ramachandran favoured, <0.2% for Ramachandran outliers, <1% for poor rotamers and 0 for C^{β} deviation outliers; other goal levels given in the summary are more arbitrary.

However, Ramachandran outliers are of course not rare in general; they increase as a function of resolution (Arendall *et al.*, 2005) and especially of B factor (Lovell *et al.*, 2003). Fig. 21.6.2.3(*a*) shows the points for an above-average 2.5 Å resolution structure plotted on the smoothed contours of the reference distribution for favoured (enclosing 98% of the good data) and allowed (enclosing 99.95% of the good data). The four Ramachandran plots are presented in kinemage and pdf form, scores are given in the multi-criterion chart (Fig. 21.6.2.2) and outliers are flagged in green on the multi-criterion kinemage (Figs. 21.6.1.1 and 21.6.1.2).

Multi-dimensional rotamer distributions for the individual side-chain types are currently contoured only down to excluding 1% of the high-quality data. [All such distributions used in *MolProbity* will be periodically updated to take advantage of the expanding reference data.] These are updates of the ‘penultimate rotamer library’



Viewing 1n78FH-multi.table

When finished, you should

Hint: Use File | Save As... to save a copy of this page.

All-Atom Contacts	Clashscore, all atoms:	25.46	24th percentile* (N=576, 2.10Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.		
Protein Geometry	Poor rotamers	6.36%	Goal: <1%
	Ramachandran outliers	0.21%	Goal: <0.2%
	Ramachandran favored	98.61%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	MolProbity score^	2.51	42nd percentile* (N=11758, 2.10Å ± 0.25Å)
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	0.64%	Goal: <0.1%
Nucleic Acid Geometry	Probably wrong sugar puckers:	15	Goal: 0
	Bad backbone conformations•:	25	Goal: 0
	Residues with bad bonds:	0.00%	Goal: 0%
	Residues with bad angles:	8.67%	Goal: <0.1%

* 100th percentile is the best among structures of comparable resolution; 0th percentile is the worst.

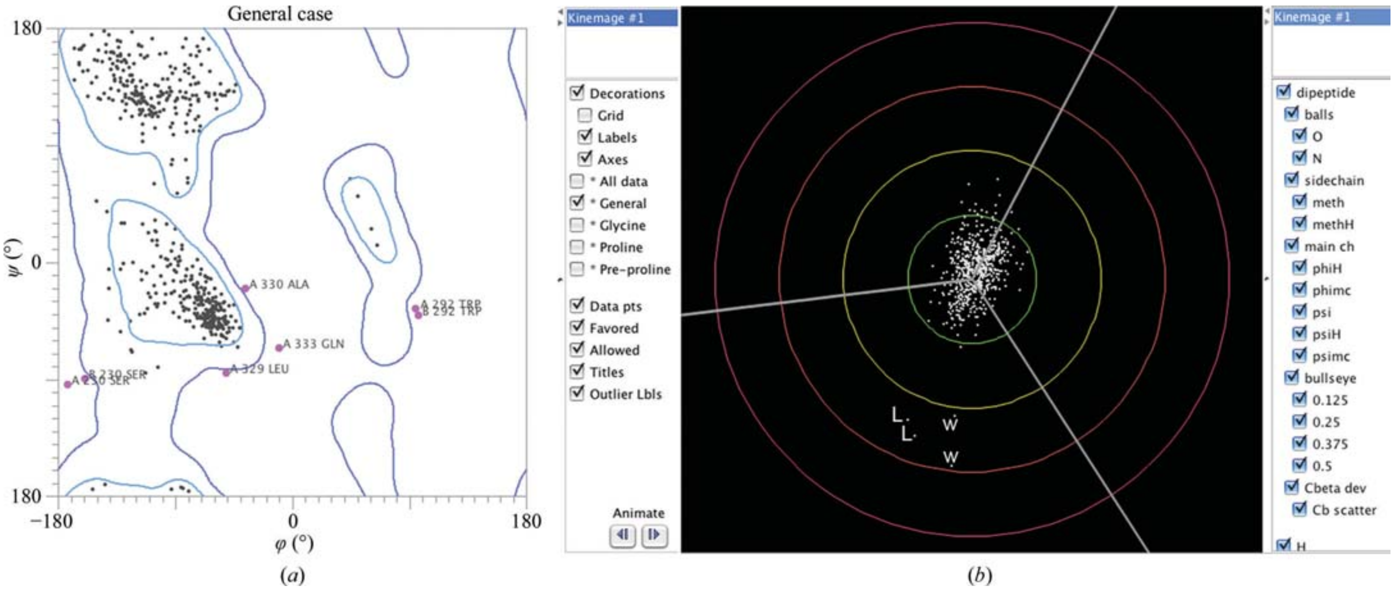
• RNA backbone was recently shown to be rotameric. Outliers are RNA suites that don't fall into recognized rotamers.

^ MolProbity score is defined as the following: 0.42574*log(1+clashscore) + 0.32996*log(1+max(0,pctRotOut-1)) + 0.24979*log(1+max(0,100-pctRamaFavored-2)) + 0.5

#	Res	High B	Clash > 0.4Å	Ramachandran	Rotamer	Cβ deviation	Base-P perp. dist.	RNA suite conf.	Bond lengths.	Bond angles.
		Avg: 35.96	Clashscore: 25.46	Outliers: 2 of 932	Poor rotamers: 50 of 786	Outliers: 0 of 868	Outliers: 15 of 150	Outliers: 25 of 150	Outliers: 0 of 1086	Outliers: 19 of 1086
D 574	C	74.81	0.599Å H5' with B 177 VAL CG2	-	-	-	suspect sugar pucker - δ & ε outlier (base-p distance indicates 2'-endo)	OUTLIER δδγ none (triaged epsilon-1)	-	2 OUTLIER(S) worst is C2'-C3'-C3': 4.876 σ
D 575	C	76.57	0.576Å H2' with D 576 A O4'	-	-	-	-	OUTLIER δδγ none (triaged epsilon-1)	-	-
D 576	A	75.23	0.795Å H5'' with B 187 TYR HB2	-	-	-	-	conformer: 1c δδγ 33 t, suiteness = 0.236	-	-
A 1	MET	59.14	0.651Å HB2 with A1556 HOH O	-	8.9% (*ptm*) chi angles: 64.6,178.7,255.3	0.014Å	-	-	-	-
A 2	VAL	33.59	-	Favored (51.36%) General case / -123.9,120.7	38.9% (*t*) chi angles: 166.1	0.025Å	-	-	-	-
A 3	VAL	23.66	-	Favored (37.26%) General case / -123.6,117.0	57.4% (*t*) chi angles: 180.5	0.041Å	-	-	-	-
A 4	THR	17.48	-	Favored (11.51%) General case / -115.5,169.1	61.1% (*p*) chi angles: 62.4	0.084Å	-	-	-	-

Figure 21.6.2.2

A *MolProbity* results summary and sortable multi-criterion chart for the 1n78 Glu tRNA-synthetase complex at 2.1 Å resolution (Sekine *et al.*, 2003). The summary gives numerical values, goals and relative percentiles for clashscore, torsion angle and geometry criteria for both protein and nucleic acid components, with traffic light colour-coding for good and bad values. Below the summary is a short extract from the detailed chart with values and specifics for each criterion on each residue. Notice that outliers (highlighted) tend to cluster.



☑ Decorations
☐ Grid
☑ Labels
☑ Axes
☐ * All data
☑ * General
☐ * Glycine
☐ * Proline
☐ * Pre-proline
☑ Data pts
☑ Favored
☑ Allowed
☑ Titles
☑ Outlier Lbls

☑ dipeptide
☑ balls
☑ O
☑ N
☑ sidechain
☑ meth
☑ methH
☑ main ch
☑ phiH
☑ phimc
☑ psi
☑ psiH
☑ psimc
☑ bullseye
☑ 0.125
☑ 0.25
☑ 0.375
☑ 0.5
☑ Cbeta dev
☑ Cb scatter
☑ H

Figure 21.6.2.3

The general case Ramachandran kinemage and the C^{β} deviation kinemage for file 2dq4. In (*a*) the φ, ψ values for each residue are plotted on a background of the smoothed contours from high-quality data (see text). Over 98% lie inside the inner ‘favoured’ 98% contour, but there are seven outliers outside the outer 99.95% contour. Gly, Pro and pre-Pro residues are on separate plots (not shown). In (*b*) the C^{β} deviation kinemage shows each residue's C^{β} position relative to an ideal C^{β} and its three bond vectors (grey lines). Circles mark the deviation distances, with the yellow circle at the 0.25 Å cutoff for outliers. Most of the distribution is good, but an adjacent Leu and Trp in each chain (labelled) are part of an outlier cluster and probably reflect distortions caused by a local fitting problem.

(Lovell *et al.*, 2000), although certainly not yet ultimate. Above the 1% level, sets of χ values are assigned to named rotamers; below 1% they are designated as poor rotamers, not outliers, since they are disfavoured but quite possible if stabilized by tight packing or a couple of good hydrogen bonds. However, there is no justification for fitting poor rotamers on the protein surface with no interactions to hold them in an unfavourable conformation. Some bad rotamers result from systematic errors arising from fitting branched side chains (Thr/Val/Ile/Leu/Arg) backwards into ambiguous density. The two major systematic errors for Leu are described in Lovell *et al.* (2000) and are given zero rotamer-quality scores in *MolProbity*; other cases are discussed in Section 21.6.3.2. Poor rotamers are scored in the multi-criterion chart and are represented by gold side chains in the multi-criterion kinemages (Figs. 21.6.1.1 and 21.6.1.2). The *MolProbity* summary (top of Fig. 21.6.2.2) reports the percentage of residues with poor rotamers, Ramachandran outliers and Ramachandran favoured conformations.

21.6.2.4. Covalent-geometry analyses

MolProbity now evaluates backbone bond-length and bond-angle outliers. Mean overall deviations in geometrical parameters are a measure of correct procedures and weights in refinement, but not of structural accuracy. Local geometry outliers $\geq 4\sigma$, however, are most often the indirect result of local misfitting and are thus very useful diagnostics, especially for bond angles. [Note that some such problems occur even at very high resolution, either at the ends of alternate conformations or in under-restrained regions with high *B* factor.] A major recent addition to *MolProbity* is the ability to assess the ideality of covalent backbone geometry for protein and RNA with the Java program *DANGLE*, which is also available for standalone command-line use from the software section at http://kinemage.biochem.duke.edu. New and more intuitive visualizations for backbone bond-length and angle deviations have been developed for display on the three-dimensional structure. For bond lengths, a 'spring' is drawn along the bond axis and scaled such that the ideal distance is equivalent to six turns and 4σ deviation corresponds to one turn of stretching or compaction (Fig. 21.6.1.1). For bond angles, bold lines are drawn to represent the ideal angle and a fan of increasingly thin lines fades out across the model *versus* ideal angle difference to highlight the extent of the deviation (Fig. 21.6.1.1). In each case model values greater (or less) than the corresponding ideal value by at least 4σ are displayed in red (or blue) and are listed on the multi-criterion chart. Protein backbone parameters were derived from Engh & Huber (Chapter 18.3) and nucleic acid backbone parameters from Parkinson *et al.* (1996).

The C^{α} atom is where local problems with backbone or side-chain fitting must be reconciled. This affects the bond angles or improper dihedrals that define the C^{β} position, but can be manifested in almost any combination of those individual parameters. Therefore, *MolProbity* evaluates the resulting overall distortion of the C^{β} position from ideality, called the C^{β} deviation (Lovell *et al.*, 2003). Residues with a C^{β} deviation of ≥ 0.25 Å are flagged in the chart and shown in the kinemage as a magenta ball centred on the ideal position (calculated from the backbone coordinates and allowing for changes in the τ angle) and tangent to the modelled position (Fig. 21.6.1.1), since we have found that values of ≥ 0.25 Å are very often correlated with some form of local misfitting. Fig. 21.6.2.3(*b*) shows a separate plot produced for all the C^{β} deviation values in a structure, shown relative to the ideal C^{β} position. The Leu and Trp C^{β} outliers in each chain form a tight turn with a suspicious peptide orientation and eight other outliers and so must represent some form of misfitting.

21.6.2.5. Nucleic acid analyses

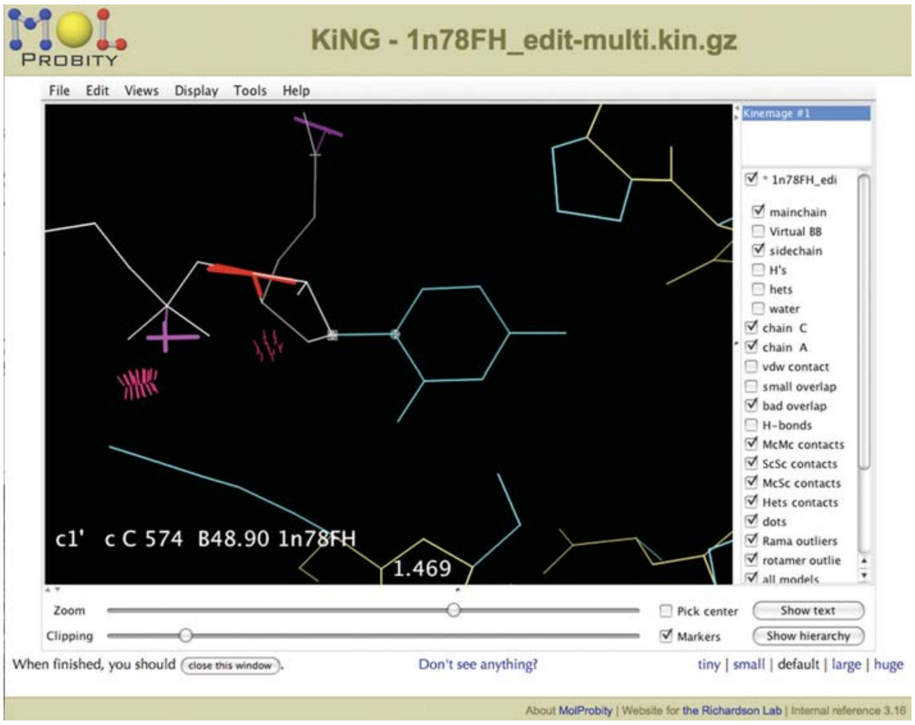
Nucleic acids are treatable equivalently to proteins for all-atom contacts and clashscore and for bond-length and bond-angle analyses, as long as the correct parameter sets are used. Both DNA and RNA show more non-uniform distribution of local problems than do proteins, with the bases and phosphates located well and the rest of the sugar-phosphate backbone very prone to errors (Word, Lovell, Richardson & Richardson, 1999; Murray *et al.*, 2003), since it has many torsion variables and rather indistinct electron density at moderate resolutions. All-atom contacts are very helpful in diagnosing backbone misfitting, especially for RNA structures, which are rapidly gaining biological interest and structural attention.

In addition, *MolProbity* now includes diagnosis of suspect ribose puckers and torsion-angle analysis of preferred RNA backbone conformers. We have found that the dominant C3′-*endo* and C2′-*endo* sugar puckers are highly correlated to the perpendicular distance between the C1′–N1/9 glycosidic bond vector and the following (3′) phosphate: >2.9 Å for C3′-*endo* and <2.9 Å for C2′-*endo*. *MolProbity* checks this distance against the modelled sugar pucker, as well as outliers in individual ε or δ values. All such outliers are listed in the multi-chart and ribose-pucker outliers are flagged in the kinemage (Fig. 21.6.1.1). An example is shown in Fig. 21.6.2.4, where what should have been a C2′-*endo* pucker (by the short perpendicular) was fitted as an intermediate unfavourable pucker close to the more common default C3′-*endo* pucker, also producing geometry and ε outliers.

High-dimensional analysis of the combinations of backbone torsion angles within an RNA 'suite' (the unit from sugar to sugar) has shown that there are distinct 'rotameric' backbone conformers. The RNA Ontology Consortium has defined a two-character nomenclature and an initial set of 54 favourable RNA backbone conformers (Richardson *et al.*, 2008). We created the *SUITENAME* program to identify either the named conformer or an outlier for each suite in an RNA structure. These conformers and their 'suiteness' quality score are listed in the *MolProbity* multi-chart.

21.6.2.6. The overall MolProbity score

In response to user demand, the '*MolProbity* score' provides a single number that represents the central *MolProbity* protein quality statistics. It is a log-weighted combination of the clashscore, percentage Ramachandran not favoured and percentage bad side-chain rotamers, giving one number that reflects the crystallographic resolution at which those values would be expected. Therefore, a structure with a numerically lower *MolProbity* score than its actual crystallographic resolution is, quality-wise, better than the average structure at that resolution. There is some distortion in the fit at very high or very low resolutions; for these ranges it is preferable to judge by the resolution-specific percentile score, which is also reported in the summary. Percentile scores are currently given for clashscore and for *MolProbity* score relative to the cohort of PDB structures within 0.25 Å of the file's resolution.


MolProbity
KiNG - 1n78FH_edit-multi.kin.gz
File Edit Views Display Tools Help
Kinemage #1
c1' c C 574 B48.90 1n78FH
1.469
Zoom
Clipping
Show text
Show hierarchy
When finished, you should close this window.
Don't see anything?
tiny | small | default | large | huge
About MolProbity | Website for the Richardson Lab | Internal reference 3.16

☑ * 1n78FH_edi
☑ mainchain
☐ Virtual BB
☑ sidechain
☐ H's
☐ hets
☐ water
☑ chain C
☑ chain A
☐ vdw contact
☐ small overlap
☑ bad overlap
☐ H-bonds
☑ McMc contacts
☑ ScSc contacts
☑ McSc contacts
☑ Hets contacts
☑ dots
☑ Rama outliers
☑ rotamer outlie
☑ all models
☐ Pick center
☑ Markers

Figure 21.6.2.4
Close-up of a ribose-pucker outlier in the multi-criterion kinemage for 1n78, with backbone and bases turned on. C574 has a short phosphate-to-glycosidic bond perpendicular (magenta line and cross), but was fitted with an intermediate pucker near C3′-*endo*. The bad pucker torques the connected groups strongly, probably causing the bond-angle outliers (red) and steric clashes (hot-pink spikes). Note that C574 is in the binding interface between RNA (white backbone) and enzyme (yellow backbone) close to the active site.

21.6.3. Correction of outliers

21.6.3.1. *Manual rebuilding*

Except for Asn/Gln/His flip corrections, *MolProbity* does not yet directly include the ability to correct the errors it finds in structures; it relies on users having access to standalone local software for rebuilding and refinement. The standalone version of *KiNG* has some rebuilding tools for modelling side chains and making small local 'backrub' adjustments to structures, with the help of electron-density display, interactive contact dots and rotamer evaluation (Davis *et al.*, 2006; Chen *et al.*, 2009; Chapter 21.5). Fig. 21.6.3.1 illustrates such a correction process in *KiNG*, rebuilding a backward-fitted leucine with a clash and a bad rotamer (one of the cases of a systematic error), resulting in an ideal geometry side chain with an excellent rotamer and well packed all-atom contacts. The top view shows that the original and rebuilt side chains fit the terminal methyls into the same rather ambiguous density, but move the C$^{\gamma}$ substantially. More recent versions of this DNA polymerase structure (*e.g.* PDB code 2hhv at 1.55 Å resolution; Warren *et al.*, 2006) all use the new conformation. Manual rebuilding is facilitated by the fact that all-atom clashes are inherently directional, as are bond-angle distortions, while a good library of rotamer choices helps the user test all the alternatives.

For more extensive refitting, a fully featured crystallographic rebuilding program such as *Coot* (Emsley *et al.*, 2010) is needed. *MolProbity* generates 'to-do' scripts that can be read into *Coot*, bringing up a button list, where each entry will zoom to a problem area. In combination with the ability of *Coot* to use *REDUCE* and *PROBE* interactively to generate all-atom contact dots, these features make it easier to address the problems diagnosed by *MolProbity*. Any rebuilding that moves atoms must of course then undergo further crystallographic refinement. Our own laboratory tested the combined cycle of *MolProbity*, rebuilding and refinement on about 30 protein structures as part of the SouthEast Collaboratory for Structural Genomics (Arendall *et al.*, 2005), finding that its early application led to a smoother structure-solution process and demonstrably better final structures. In addition to backward-fitted side chains, commonly corrected problems included peptide flips, switched backbone and side chain near chain ends, 'waters' that were really ions, noise peaks or unfit alternate conformations and occasionally a shift in sequence register. Many other crystallographic groups have since adopted these methods.

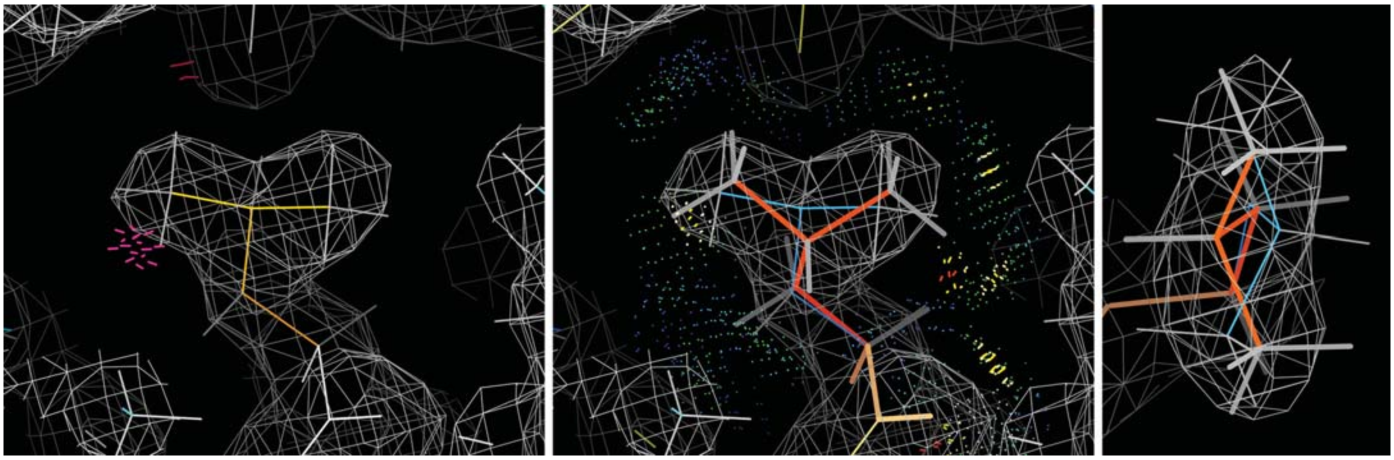


Figure 21.6.3.1
Rebuilding of a backward-fitted Leu side chain in *KiNG* off-line in the DNA polymerase 1xwl at 1.7 Å resolution (Kiefer *et al.*, 1997). The original (left) fits the density fairly well but is a rotamer outlier with a clash. One of the two best Leu rotamers also fits the density well with good all-atom packing. The top view (right) shows the 180° relationship of the two conformations.

21.6.3.2. Automated corrections

For correcting RNA-suite outliers, we have collaboratively developed the independent program *RNABC* (Wang *et al.*, 2008), which performs an automated search for more suitable backbone conformations of an RNA suite diagnosed with a bad ribose pucker or serious clashes. It leaves the more accurately determined bases and P atoms fixed in place and performs a pruned but systematic search through the other parameters, outputting all acceptable alternatives found within user-set tolerance limits.

Recently, we have developed and tested the *AUTOFIX* program for automated correction of diagnosed backward-fitted Thr, Val, Leu and Arg side chains (Headd *et al.*, 2009). In contrast to Asn/Gln/His flips, which simply exchange atoms and do not change the agreement with the data, these more complex side chains require real-space refinement in order to determine the proper correction and crystallographic re-refinement after the approximate 180° flips have been made. The original version used *Coot* to perform rotamer selection and real-space refinement for the proposed corrections, with *MolProbity* diagnosis before and after. Results were checked by re-refinement. Run on a sample of 945 PDB files, *AUTOFIX* accepted corrections for over 40% of diagnosed bad Thr, Val and Leu side chains and 15% of bad Arg side chains, or 3679 corrected side chains. The most important of our requirements for *AUTOFIX* is that it does no harm; we are willing to miss some of the possible corrections in order to ensure that those we accept are essentially always true improvements. Thr and Arg, in particular, make hydrogen bonds that are often important at active sites or binding interfaces and since they are asymmetrical these interactions change drastically if the side chain is fitted backwards. Such improvements were often seen in the test set. A successor to *AUTOFIX* has been built into *phenix.refine*, called by setting fix_rotamers = True.

21.6.4. Other *MolProbity* utility functions

21.6.4.1. Interface analysis

PROBE can also be used to calculate the all-atom contacts at interfaces, *e.g.* between two chains of a structure or between a protein and a ligand. Access to this feature is provided in *MolProbity* by the 'Visualize interface contacts' analysis option after H atoms have been added. The user is required to choose the chains and/or the molecular types for which to calculate the contacts (*e.g.* protein *versus* protein or protein *versus* heteroatoms or RNA). This functionality creates both a kinemage with the resulting all-atom contacts displayed on the model and a text list of the atom pairs in contact.

21.6.4.2. Protein loop fitting

MolProbity includes the Java software *JIFFILOOP* (Chen, 2010) for providing potential protein-fragment conformations that can fit within a gap in a protein structure. We have defined a seven-parameter system that describes the spatial relationship between any two peptides. Briefly, this system consists of the sequence separation, the distance between the two inner C$^{\alpha}$ atoms, two pseudo-angles and three pseudo-dihedrals. We used this system to create a library of *B*-factor-filtered fragments from one to 15 peptides long from our Top5200 database of structures, a set of structures chosen from each 70% nonredundant group defined by the PDB, requiring resolution < 2.0 while optimizing *MolProbity* score. *MolProbity* runs *JIFFILOOP* to search this library for candidate fragments to fill gaps within a structure. Alternatively, users can enter beginning and ending residue numbers and *MolProbity* will search for fragments which can fit between those two residues. Because this process can be fairly time-intensive, *JIFFILOOP* is not listed under 'Suggested Tools' and is currently only accessible under 'All Tools' or at the Site map. Also, owing to the size of this package, it must be added separately to the installation for a standalone *MolProbity* server.

21.6.4.3. Kinemage construction and viewing

MolProbity provides scripts (under the 'Make simple kinemages' option) for constructing a number of commonly used kinemage three-dimensional interactive visualization options such as ribbons and various types of stick figures. This functionality is useful for quick browsing of a structure or for initial creation of an illustration or presentation. The file-input page can also accept upload of pre-existing kinemage files for direct on-line viewing within the built-in kinemage viewer *KiNG*.

21.6.4.4. Other file types and functions

MolProbity uses a built-in PDB 'het_dictionary' for the information needed to add H atoms to small-molecule ligands. The user can construct and read in a custom dictionary if their file contains novel ligands. There is also provision for either uploading or fetching an electron-density map from the Electron Density Server (Kleywegt *et al.*, 2004) in any of several formats to view on-line in *KiNG* together with the model and validation results. To investigate functional sites that span across asymmetric units, one can fetch a biological unit file from the PDB. In the file-editing feature, the user can specify whether multiple 'models' are alternatives (as in an NMR ensemble) or have been pressed into service for the extra chains in the biological unit. Some X-ray structures are now treated as ensembles. For such cases, *MolProbity* internally splits the models and analyses them separately, but constructs an outlier summary strip-chart and a multi-model multi-criterion validation kinemage with both the models and their features under on/off button control. File editing also allows the deletion of chains either before or after hydrogen placement, specifying the resolution of the structure if not given in the file header or removing unwanted H atoms. These tools make it easier and faster to analyse particular parts of a structure using *MolProbity* and they help to maintain compatibility with other older software. These options are always available as separate utility functions, independent of validation or hydrogen content.

21.6.4.5. PDB-format interconversion

The release of the remediated PDB version 3.0 format in August 2007 included a number of significant changes, particularly to H-atom names and to nucleic acid residue and atom names. In order to maintain compatibility with the PDB, we converted the entire *MolProbity* core to use the new format by default. This included updating *REDUCE*, *PROBE*, *KiNG* and *PREKIN*. However, we realized that users might need to analyse files that were still in the older PDB version 2.3 format. In order to maintain backwards compatibility, we created a Remediator script (available as a standalone Perl or Python script) that can interconvert between the old and the new PDB formats. Whenever a file is input, *MolProbity* will scan for the presence of old-format atom names and, if it detects any, then it will run the Remediator script to automatically convert the input file to the new format. After analysis there is then an option available to run the Remediator script and downgrade the output file back into the old version 2.3 format if needed. This allows use of the

MolProbity analysis tools even together with older software that has not been updated to use the new format.

21.6.5. Discussion

21.6.5.1. Global versus local, absolute versus comparative

There are three quite different purposes served by structure validation: a gatekeeper function on quality for reviewers or organizations, an aid to crystallographers for obtaining the most model accuracy from their data, and a guide to end users for choosing appropriate structures and confidence levels for the conclusions they want to draw.

Validation criteria also come in distinct flavours. Those based on the diffraction data are inherently global with respect to the model; for instance, resolution (which is still the most valuable single-factor estimate of model accuracy) and $R_{\rm free}$ (Brünger, 1992). On the data side, there are also gatekeeper checks for unusual problems such as twinning or gross data incompleteness. R.m.s.d. or r.m.s.*Z*. of deviations from geometrical target values are global, but they only evaluate procedural aspects of refinement and have little to do with model accuracy. Most other validation criteria are inherently local (at the residue or even atom level), including *B* factor, real-space measures such as RSR-*Z* (Kleywegt *et al.*, 2004) and model-only measures such as the various *MolProbity* criteria described here. Any local measure becomes global when expressed in some normalized form across the entire structure, such as an average, a distribution match or a percentage occurrence of outliers.

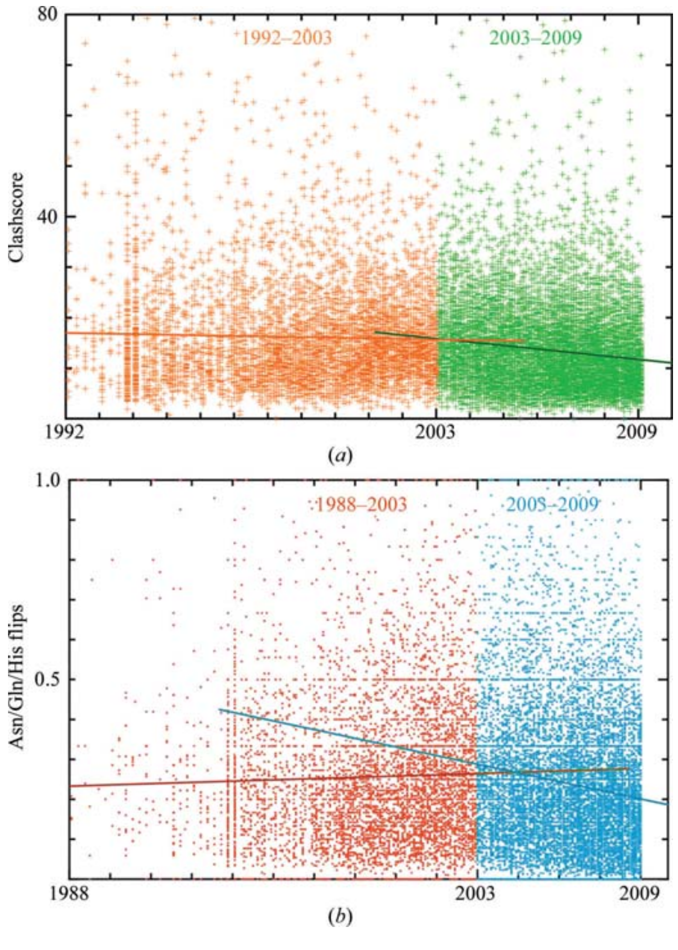


Figure 21.6.5.1
MolProbity-relevant quality criteria as a function of time for all structures in the PDB at a middle range of resolution, separately fitted before and after introduction of the web site. (*a*) All-atom clashscore (see Section 21.6.2.2); (*b*) percentage of Asn/Gln/His flips (see Section 21.6.2.1).

Strictly local measures are usually not resolution-dependent, but their globally defined versions often are. For some purposes, the desirable form of measure is a comparison (usually a percentile rank) with the cohort of PDB structures at similar resolution. *MolProbity* currently provides resolution-group percentiles for clashscore and for *MolProbity* score and will probably expand that to other criteria.

Reviewers/gatekeepers are primarily interested in global relative measures such as resolution-dependent percentiles and to some extent in absolute local flags for judging the support behind specific claims. Crystallographers need global relative measures to judge how well they have made use of their data, but it is the local measures, especially specific outliers, that are crucial to helping them to achieve a more accurate structure and to avoid making any dubious claims in poor local regions (such as an invisible inhibitor). End users need absolute global measures to choose between structures and absolute local measures to judge the reliability of the particular features they find of interest.

Because of the importance of improving and evaluating the accuracy of individual details of biological importance, both in each structure and in the database as a whole, we have chosen in *MolProbity* to emphasize calculation and user-friendly display of local indicators. We have also tried to minimize ‘false alarms’, so that a flagged outlier is almost always worth a close look.

21.6.5.2. Impact on database quality

Since *MolProbity* was first made available in late 2002, serious user work sessions (performing some operation on an input coordinate file) have multiplied by a large factor each year, with a cumulative total that is now approaching 100 000 by thousands of distinct users. In addition, many companies and structural genomics centres run their own *MolProbity* servers internally and some aspects have been incorporated into other software or meta-servers. 80% of *MolProbity* input files are uploaded, presumably by working structural biologists, and the rest are fetched from databases, presumably by end users. Those end users also include students, since *MolProbity* is increasingly being used for instructional exercises in biochemistry classes from high school to graduate level.

MolProbity’s unique feature is clash analysis from all-atom contacts, which provides sensitive new evaluation independent of refinement targets. Not surprisingly, the average clashscore remained constant (either globally or by resolution) up to 2002, since there was then no feasible way of targeting or even measuring all-atom clashes. The percentage of incorrect Asn/Gln/His flips also remained level or rose slightly prior to 2003, despite the availability of a hydrogen-bond-based system in *WHAT IF* (Hooft *et al.*, 1996), and even while refinement methods, automation and Ramachandran and rotamer quality all improved.

To evaluate the contribution *MolProbity* has made to crystallographic model quality in general, we have therefore plotted clashscore and Asn/Gln/His flips as a function of time in Fig. 21.6.5.1, with separate linear fits before and after the end of 2002. Gratifyingly, in both cases there is a clear trend of improvement since 2003. Median values also improve very steadily over that period. Anecdotal evidence indicates that this trend is mainly a consequence of thorough adoption of *MolProbity*-based methods by a small but growing fraction of crystallographers and there is therefore still much scope for further improvement in the future.

21.6.6. *MolProbity* availability

MolProbity is freely available for download from http://molprobity.biochem.duke.edu for use as a local server. This option requires either Linux or MacOSX, together with PHP and Apache. Instructions for installing *MolProbity* locally are included with the download. Having a local installation allows users to access the *MolProbity* analysis tools without internet access, as well as allowing companies with privacy or confidentiality concerns to use *MolProbity*. However, one of the most significant advantages to having a local installation of *MolProbity* is access to command-line tools. These tools provide access to the major analysis tools in *MolProbity* without having to use the web interface. Also, several scripts are included which allow users to run *MolProbity* analysis on a set of files rather than just one at a time. Some of the more useful command-line scripts include the following: scripts for adding H atoms, with or without flips, a script for obtaining overall scores for a set of files and a script for calculating a residue-by-residue analysis of a structure.

For users of the *PHENIX* crystallography system (Adams *et al.*, 2010), nearly all of the main *MolProbity* quality-analysis tools have been incorporated directly into *PHENIX* and are accessible through command-line tools or in the *PHENIX* GUI, including *REDUCE*, *PROBE*, *RAMALYZE*, *ROTALYZE*, *CBETADEV*, *RNA_VALIDATE* and *CLASHSCORE*. Both tabular and graphical results are provided, including three-dimensional display in *KiNG* and links to individual outliers in *Coot*.

All of the individual programs called by *MolProbity* are also available, multi-platform and open source, from the software section at http://kinemage.biochem.duke.edu.

Ian Davis is the initial developer of the *MolProbity* site, Xueyi Wang and Jack Snoeyink programmed *RNABC*, Jeremy Block developed *MolProbity* functions for NMR and the *PHENIX* development team at LBL were very helpful with *AUTOFIX* and with integration of *MolProbity* functions. *MolProbity* development is supported by NIH grant GM73919.

This chapter is based a recent article published in *Acta Crystallographica Section D* (Chen *et al.*, 2010).

References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *PHENIX: a comprehensive Python-based system for macromolecular structure solution*. *Acta Cryst.* D**66**, 213–221.

Arendall, W. B. III, Tempel, W., Richardson, J. S., Zhou, W., Wang, S., Davis, I. W., Liu, Z.-J., Rose, J. P., Carson, W. M., Luo, M., Richardson, D. C. & Wang, B.-C. (2005). *A test of enhancing model accuracy in high-throughput crystallography*. *J. Struct. Funct. Genomics*, **6**, 1–11.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Brünger, A. T. (1992). *Free R-value – a novel statistical quantity for assessing the accuracy of crystal structures*. *Nature (London)*, **355**, 472–475.

Chen, V. B. (2010). *Building better backbones: visualizations, analyses, and tools for higher quality macromolecular structure models*. PhD thesis, Duke University, USA.

Chen, V. B., Arendall, W. B. III, Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *MolProbity: all-atom structure validation for macromolecular crystallography*. *Acta Cryst.* D**66**, 12–21.

Chen, V. B., Davis, I. W. & Richardson, D. C. (2009). *KiNG (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program*. *Protein Sci.* **18**, 2403–2409.

Davis, I. W., Arendall, W. B. III, Richardson, D. C. & Richardson, J. S. (2006). *The backrub motion: how protein backbone shrugs when a sidechain dances*. *Structure*, **14**, 265–274.

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids*. *Nucleic Acids Res.* **35**, W375–W383.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Features and development of Coot*. *Acta Cryst.* D**66**, 486–501.

Headd, J. J., Immormino, R. M., Keedy, D. A., Emsley, P., Richardson, D. C. & Richardson, J. S. (2009). *Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place*. *J. Struct. Funct. Genomics*, **10**, 83–93.

Higman, V. A., Boyd, J., Smith, L. J. & Redfield, C. (2004). *Asparagine and glutamine side-chain conformation in solution and crystal: A comparison for hen egg-white lysozyme using residual dipolar couplings*. *J. Biomol. NMR*, **30**, 327–346.

Hooft, R. W. W., Sander, C. & Vriend, G. (1996). *Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures*. *Proteins*, **26**, 363–376.

Kiefer, J. R., Mao, C., Hansen, C. J., Basehore, S. L., Hogrefe, H. H., Braman, J. C. & Beese, L. S. (1997). *Crystal structure of a thermostable Bacillus DNA polymerase I large fragment at 2.1 Å resolution*. *Structure*, **5**, 95–108.

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *The Uppsala Electron-Density Server*. *Acta Cryst.* D**60**, 2240–2249.

Li, S., Finley, J., Liu, Z.-J., Qiu, S.-H., Chen, H., Luan, C.-H., Carson, M., Tsao, J., Johnson, D., Lin, G., Zhao, J., Thomas, W., Nagy, L. A., Sha, B., DeLucas, L. J., Wang, B.-C. & Luo, M. (2002). *Crystal structure of the cytoskeleton-associated protein glycine-rich (CAP-Gly) domain*. *J. Biol. Chem.* **277**, 48596–48601.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Structure validation by Cα geometry: φ, ψ and Cβ deviation*. *Proteins*, **50**, 437–450.

Lovell, S. C., Word, J. M., Richardson, J. S. & Richardson, D. C. (2000).*The penultimate rotamer library*. *Proteins*, **40**, 389–408.

Murray, L. J. W., Arendall, W. B. III, Richardson, D. C. & Richardson, J. S. (2003). *RNA backbone is rotameric*. *Proc. Natl Acad. Sci. USA*, **100**, 13904–13909.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures*. *Acta Cryst.* D**52**, 57–64.

Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkovits, E., Williams, L. D., Keating, K. S., Pyle, A. M., Micallef, D., Westbrook, J. & Berman, H. M. (2008). *RNA backbone: consensus all-angle conformers and modular string nomenclature (an RNA Ontology Consortium contribution)*. *RNA*, **14**, 465–481.

Sekine, S., Nureki, O., Dubois, D. Y., Bernier, S., Chenevert, R., Lapointe, J., Vassylyev, D. G. & Yokoyama, S. (2003). *ATP binding by glutamyl-tRNA synthetase is switched to the productive mode by tRNA binding*. *EMBO J.* **22**, 676–688.

Wang, X., Kapral, G., Murray, L., Richardson, D., Richardson, J. & Snoeyink, J. (2008). *RNABC: forward kinematics to reduce all-atom steric clashes in RNA backbone*. *J. Math. Biol.* **56**, 253–278.

Warren, J. J., Forsberg, L. J. & Beese, L. S. (2006). *The structural basis for the mutagenicity of O(6)-methyl-guanine lesions*. *Proc. Natl Acad. Sci. USA*, **103**, 19701–19706.

Word, J. M., Lovell, S. C., LaBean, T. H., Taylor, H. C., Zalis, M. E., Presley, B. K., Richardson, J. S. & Richardson, D. C. (1999). *Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms*. *J. Mol. Biol.* **285**, 1711–1733.

Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. (1999). *Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation*. *J. Mol. Biol.* **285**, 1735–1747.

International Tables for Crystallography (2012). Vol. F, Chapter 22.1, pp. 703–712.

PART 22. MOLECULAR GEOMETRY AND FEATURES

Chapter 22.1. Protein geometry: volumes, areas and distances

M. Gerstein and F. M. Richards

22.1.1. Introduction

For geometric analysis, a protein consists of a set of points in three dimensions. This information corresponds to the actual data provided by the experiment, which are fundamentally of a geometric rather than chemical nature. That is, crystallography primarily tells one about the positions of atoms and perhaps an approximate atomic number, but not their charge or number of hydrogen bonds.

For the purposes of geometric calculation, each point has an assigned identification number and a position defined by three coordinates in a right-handed Cartesian system. (These coordinates will be based on the electron density for X-ray derived structures and on nuclear positions for those derived from neutron scattering. Each coordinate is usually assumed to have an accuracy between 0.5 and 1.0 Å.) Normally, only one additional characteristic is associated with each point: its size, usually measured by a van der Waals (VDW) radius. Furthermore, characteristics such as chemical nature and covalent connectivity, if needed, can be obtained from lookup tables keyed on the ID number.

Our model of a protein, thus, is the van der Waals envelope – the set of interlocking spheres drawn around each atomic centre. In brief, the geometric quantities of the model of particular concern in this section are its total surface area, total volume, the division of these totals among the amino-acid residues and individual atoms, and the description of the empty space (cavities) outside the van der Waals envelope. These values are then used in the analysis of protein structure and properties.

All the geometric properties of a protein (*e.g.* surfaces, volumes, distances *etc.*) are obviously interrelated. So the definition of one quantity, *e.g.* area, obviously impacts on how another, *e.g.* volume, can be consistently defined. Here, we will endeavour to present definitions for measuring protein volume, showing how they are related to various definitions of linear distance (VDW parameters) and surface. Further information related to macromolecular geometry, focusing on volumes, is available from http://www.molmovdb.org/geometry/.

22.1.2. Definitions of protein volume

22.1.2.1. Volume in terms of Voronoi polyhedra: overview

Protein volume can be defined in a straightforward sense through a particular geometric construction called the Voronoi polyhedron. In essence, this construction provides a useful way of partitioning space amongst a collection of atoms. Each atom is surrounded by a single convex polyhedron and allocated the space within it (Fig. 22.1.2.1). The faces of Voronoi polyhedra are formed by constructing dividing planes perpendicular to vectors connecting atoms, and the edges of the polyhedra result from the intersection of these planes.

Voronoi polyhedra were originally developed by Voronoi (1908) nearly a century ago. Bernal & Finney (1967) used them to study the structure of liquids in the 1960s. However, despite the general utility of these polyhedra, their application to proteins was limited by a serious methodological difficulty. While the Voronoi construction is based on partitioning space amongst a collection of 'equal' points, all protein atoms are not equal. Some are clearly larger than others. In 1974, a solution was found to this problem (Richards, 1974), and since then Voronoi polyhedra have been applied to proteins.

22.1.2.2. The basic Voronoi construction

22.1.2.2.1. Integrating on a grid

The simplest method for calculating volumes with Voronoi polyhedra is to put all atoms in the system on a fine grid. Then go to each grid point (*i.e.* voxel) and add its infinitesimal volume to the atom centre closest to it. This is prohibitively slow for a real

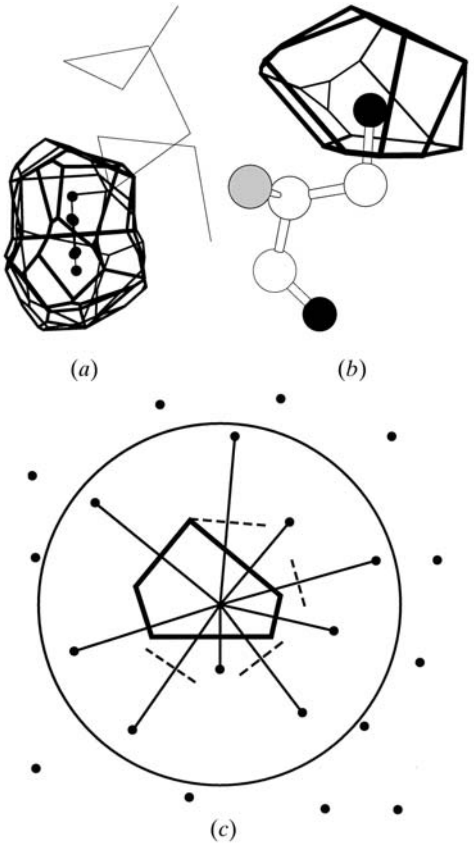


Figure 22.1.2.1
The Voronoi construction in two and three dimensions. Representative Voronoi polyhedra from 1CSE (subtilisin) are shown. (*a*) Six polyhedra around the atoms in a Phe ring. (*b*) A single polyhedron around the side-chain hydroxyl oxygen (OG) of a serine. (*c*) A schematic showing the construction of a Voronoi polyhedron in two dimensions. The broken lines indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.2.4).

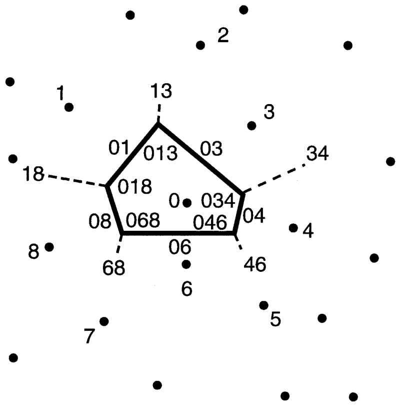


Figure 22.1.2.2
Labelling parts of Voronoi polyhedra. The central atom is atom 0, and each neighbouring atom has a sequential index number (1, 2, 3 . . .). Consequently, in three dimensions, planes are denoted by the indices of the two atoms that form them (*e.g.* 01); lines are denoted by the indices of three atoms (*e.g.* 012); and vertices are denoted by four indices (*e.g.* 0123). In the two-dimensional representation shown here, lines are denoted by two indices, and vertices by three. From a collection of points, a volume can be calculated by a variety of approaches: First of all, the volume of a tetrahedron determined by four points can be calculated by placing one vertex at the origin and evaluating the determinant formed from the remaining three vertices. (The tetrahedron volume is one-sixth of the determinant value.) The determinant can be quickly calculated by a vector triple product, $\mathbf{w} \cdot (\mathbf{u} \times \mathbf{v})$, where $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are vectors between the vertex selected to be the origin and the other three vertices of the tetrahedron. Alternatively, the volume of the pyramid from a central atom to a face can be calculated from the usual formula $Ad/3$, where A is the area of the face and d is the distance to the face.

protein structure, but it can be made somewhat faster by randomly sampling grid points. It is, furthermore, a useful approach for high-dimensional integration (Sibbald & Argos, 1990).

More realistic approaches to calculating Voronoi volumes have two parts: (1) for each atom find the vertices of the polyhedron around it and (2) systematically collect these vertices to draw the polyhedron and calculate its volume.

22.1.2.2.2. *Finding polyhedron vertices*

In the basic Voronoi construction (Fig. 22.1.2.1), each atom is surrounded by a unique limiting polyhedron such that all points within an atom's polyhedron are closer to this atom than all other atoms. Consequently, points equidistant from two atoms lie on a dividing plane; those equidistant from three atoms are on a line, and those equidistant from four atoms form a vertex. One can use this last fact to find all the vertices associated with an atom easily. With the coordinates of four atoms, it is straightforward to solve for possible vertex coordinates using the equation of a sphere. [That is, one uses four sets of coordinates (x, y, z) and the equation $(x-a)^2 + (y-b)^2 + (z-c)^2 = r^2$ to solve for the centre (a, b, c) and radius (r) of the sphere.] One then checks whether this putative vertex is closer to these four atoms than any other atom; if so, it is a real vertex.

Note that this procedure can fail for certain pathological arrangements of atoms that would not normally be encountered in a real protein structure. These occur if there is a centre of symmetry, as in a regular cubic lattice or in a perfect hexagonal ring in a protein (see Procacci & Scateni, 1992). Centres of symmetry can be handled (in a limited way) by randomly perturbing the atoms a small amount and breaking the symmetry. Alternatively, the 'chopping-down' method described below is not affected by symmetry centres – an important advantage to this method of calculation.

22.1.2.2.3. *Collecting vertices and calculating volumes*

To collect the vertices associated with an atom systematically, label each one by the indices of the four atoms with which it is associated (Fig. 22.1.2.2). To traverse the vertices on one face of a polyhedron, find all vertices that share two indices and thus have two atoms in common, *e.g.* a central atom (atom 0) and another atom (atom 1). Arbitrarily pick a vertex to start at and walk around the perimeter of the face. One can tell which vertices are connected by edges because they will have a third atom in common (in addition to atom 0 and atom 1). This sequential walking procedure also provides a way of drawing polyhedra on a graphics device. More importantly, with reference to the starting vertex, the face can be divided into triangles, for which it is trivial to calculate areas and volumes (see Fig. 22.1.2.2 for specifics).

22.1.2.3. *Adapting Voronoi polyhedra to proteins*

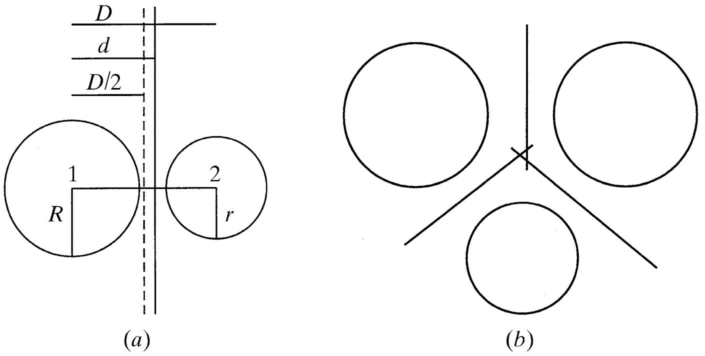
In the procedure outlined above, all atoms are considered equal, and the dividing planes are positioned midway between atoms (Fig. 22.1.2.3). This method of partition, called bisection, is not physically reasonable for proteins, which have atoms of obviously different size (such as oxygen and sulfur). It chemically misallocates volume, giving excess to the smaller atom.

Two principal methods of repositioning the dividing plane have been proposed to make the partition more physically reasonable: method B (Richards, 1974) and the radical-plane method (Gellatly & Finney, 1982). Both methods depend on the radii of the atoms in contact (R for the larger atom and r for the smaller one) and the distance between the atoms (D). As shown in Fig. 22.1.2.3, they position the plane at a distance d from the larger atom. This distance is always set such that the plane is closer to the smaller atom.

22.1.2.3.1. *Method B and a simplification of it: the ratio method*

Method B is the more chemically reasonable of the two and will be emphasized here. For atoms that are covalently bonded, it divides the distance between the atoms proportionaly according to their covalent-bond radii:

$$d = DR/(R + r). \qquad (22.1.2.1)$$



(*a*) (*b*)

Figure 22.1.2.3
Positioning of the dividing plane. (*a*) The dividing plane is positioned at a distance d from the larger atom with respect to radii of the larger atom (R) and the smaller atom (r) and the total separation between the atoms (D). (*b*) Vertex error. One problem with using method B is that the calculation does not account for all space, and tiny tetrahedra of unallocated volume are created near the vertices of each polyhedron. Such an error tetrahedron is shown. The radical-plane method does not suffer from vertex error, but it is not as chemically reasonable as method B.

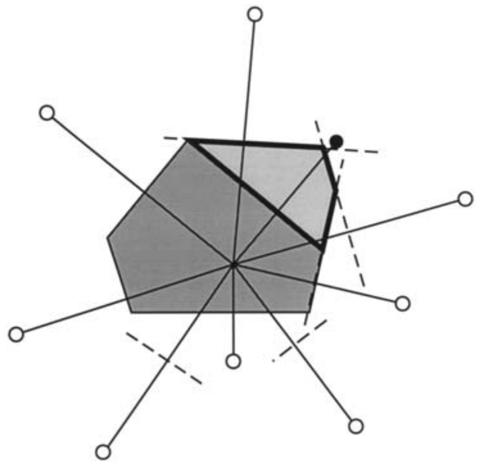


Figure 22.1.2.4
The 'chopping-down' method of polyhedra construction. This is necessary when using method B for plane positioning, since one can no longer solve for the position of vertices. One starts with a large tetrahedron around the central atom and then 'chops it down' by removing vertices that are outside the plane formed by each neighbour. For instance, say vertex 0214 of the current polyhedron is outside the plane formed by neighbour 6. One needs to delete 0214 from the list of vertices and recompute the polyhedron using the new vertices formed from the intersection of the plane formed by neighbour 6 and the current polyhedron. Using the labelling conventions in Fig. 22.1.2.2, one finds that these new vertices are formed by the intersection of three lines (021, 024 and 014) with plane 06. Therefore one adds the new vertices 0216, 0246 and 0146 to the polyhedron. However, there is a snag: it is necessary to check whether any of the three lines are not also outside of the plane. To do this, when a vertex is deleted, all the lines forming it (*e.g.* 021, 024, 014) are pushed onto a secondary list. Then when another vertex is deleted, one checks whether any of its lines have already been deleted. If so, this line is not used to intersect with the new plane. This process is shown schematically in two dimensions. For the purposes of the calculations, it is useful to define a plane created by a vector $\mathbf{v}$ from the central atom to the neighbouring atom using a constant K so that for any point $\mathbf{u}$ on the plane $\mathbf{u}\cdot\mathbf{v} = K$. If $\mathbf{u}\cdot\mathbf{V} > K$, $\mathbf{u}$ is on the wrong side of the plane, otherwise it is on the right side. A vertex point $\mathbf{w}$ satisfies the equations of three planes: $\mathbf{w}\cdot\mathbf{v}_1 = K_1$, $\mathbf{w}\cdot\mathbf{v}_2 = K_2$ and $\mathbf{w}\cdot\mathbf{v}_3 = K_3$. These three equations can be solved to give the components of $\mathbf{w}$. For example, the x component is given by

$$w_x = \begin{pmatrix} K_1 & v_{1y} & v_{1z} \\ K_2 & v_{2y} & v_{2z} \\ K_3 & v_{3y} & v_{3z} \end{pmatrix} \Bigg/ \begin{pmatrix} v_{1x} & v_{1y} & v_{1z} \\ v_{2x} & v_{2y} & v_{2z} \\ v_{3x} & v_{3y} & v_{3z} \end{pmatrix}.$$

For atoms that are not covalently bonded, method B splits the remaining distance between them after subtracting their VDW radii:

$$d = R + (D - R - r)/2. \qquad (22.1.2.2)$$

For separations that are not very different to the sum of the radii, the two formulae for method B give essentially the same result. Consequently, it is worthwhile to try a slight simplification of method B, which we call the 'ratio method'. Instead of using equation (22.1.2.1) for bonded atoms and equation (22.1.2.2) for non-bonded ones, one can just use equation (22.1.2.2) in both cases with either VDW or covalent radii (Tsai *et al.*, 2001). Doing this gives more consistent reference volumes (manifest in terms of smaller standard deviations about the mean).

22.1.2.3.2. Vertex error

If bisection is not used to position the dividing plane, it is much more complicated to find the vertices of the polyhedron, since a vertex is no longer equidistant from four atoms. Moreover, it is also necessary to have a reasonable scheme for 'typing' atoms and assigning them radii.

More subtly, when using the plane positioning determined by method B, the allocation of space is no longer mathematically perfect, since the volume in a tiny tetrahedron near each polyhedron vertex is not allocated to any atom (Fig. 22.1.2.3). This is called vertex error. However, calculations on periodic systems have shown that, in practice, vertex error does not amount to more than 1 part in 500 (Gerstein *et al.*, 1995).

22.1.2.3.3. 'Chopping-down' method of finding vertices

Because of vertex error and the complexities in locating vertices, a different algorithm has to be used for volume calculation with method B. (It can also be used with bisection.) First, surround the central atom (for which a volume is being calculated) by a very large, arbitrarily positioned tetrahedron. This is initially the 'current polyhedron'. Next, sort all neighbouring atoms by distance from the central atom and go through them from nearest to farthest. For each neighbour, position a plane perpendicular to the vector connecting it to the central atom according to the predefined proportion (*i.e.* from the method B formulae or bisection). Since a Voronoi polyhedron is always convex, if any vertices of the current polyhedron are on the other side of this plane to the central atom, they cannot be part of the final polyhedron and should be discarded. After this has been done, the current polyhedron is recomputed using the plane to 'chop it down'. This process is shown schematically in Fig. 22.1.2.4. When it is finished, one has a list of vertices that can be traversed to calculate volumes, as in the basic Voronoi procedure.

22.1.2.3.4. Radical-plane method

The radical-plane method does not suffer from vertex error. In this method, the plane is positioned according to

$$d = (D^2 + R^2 - r^2)/2D. \qquad (22.1.2.3)$$

22.1.2.4. Delaunay triangulation

Voronoi polyhedra are closely related (*i.e.* dual) to another useful geometric construction called the Delaunay triangulation. This consists of lines, perpendicular to Voronoi faces, connecting each pair of atoms that share a face (Fig. 22.1.2.5).

Delaunay triangulation is described here as a derivative of the Voronoi construction. However, it can be constructed directly from the atom coordinates. In two dimensions, one connects with a triangle any triplet of atoms if a circle through them does not enclose any additional atoms. Likewise, in three dimensions one connects four atoms with a tetrahedron if the sphere through them does not contain any further atoms. Notice how this construction is equivalent to the specification for Voronoi polyhedra and, in a sense, is simpler. One can immediately see the relationship between the triangulation and the Voronoi volume by noting that the volume is the distance between neighbours (as determined by the triangulation) weighted by the area of each polyhedral face. In practice, it is often easier in drawing to construct the triangles first and then build the Voronoi polyhedra from them.

Delaunay triangulation is useful in many 'nearest-neighbour' problems in computational geometry, *e.g.* trying to find the neighbour of a query point or finding the largest empty circle in a

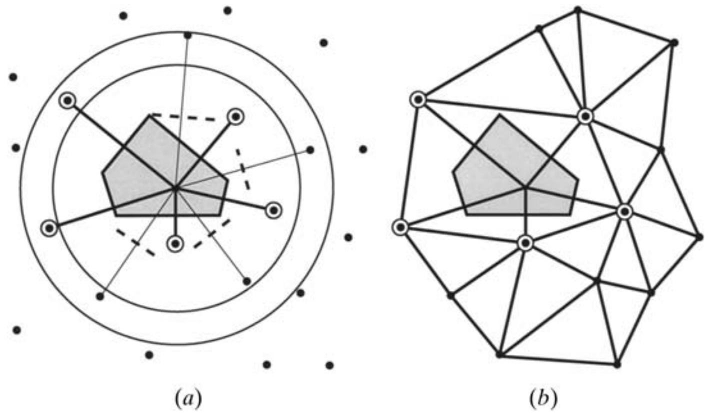


Figure 22.1.2.5
Delaunay triangulation and its relation to the Voronoi construction. (*a*) A standard schematic of the Voronoi construction. The atoms used to define the Voronoi planes around the central atom are circled. Lines connecting these atoms to the central one are part of the Delaunay triangulation, which is shown in (*b*). Note that atoms included in the triangulation cannot be selected strictly on the basis of a simple distance criterion relative to the central atom. The two circles about the central atoms illustrate this. Some atoms within the outer circle but outside the inner circle are included in the triangulation, but others are not. In the context of protein structure, Delaunay triangulation is useful in identifying true 'packing contacts', in contrast to those contacts found purely by distance threshold. The broken lines in (*a*) indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.2.4).

collection of points (O'Rourke, 1994). Since this triangulation has the 'fattest' possible triangles, it is the choice for procedures such as finite-element analysis.

In terms of protein structure, Delaunay triangulation is the natural way to determine packing neighbours, either in protein structure or molecular simulation (Singh *et al.*, 1996; Tsai *et al.*, 1996, 1997). Its advantage is that the definition of a neighbour does not depend on distance. The alpha shape is a further generalization of Delaunay triangulation that has proven useful in identifying ligand-binding sites (Edelsbrunner *et al.*, 1996, 1995; Edelsbrunner & Mucke, 1994; Peters *et al.*, 1996).

22.1.3. Definitions of protein surface

22.1.3.1. The problem of the protein surface

When one is carrying out the Voronoi procedure, if a particular atom does not have enough neighbours the 'polyhedron' formed around it will not be closed, but rather will have an open, concave shape. As it is not often possible to place enough water molecules in an X-ray crystal structure to cover all the surface atoms, these 'open polyhedra' occur frequently on the protein surface (Fig. 22.1.3.1). Furthermore, even when it is possible to define a closed polyhedron on the surface, it will often be distended and too large. This is the problem of the protein surface in relation to the Voronoi construction.

There are a number of practical techniques for dealing with this problem. First, one can use very high resolution protein crystal structures, which have many solvent atoms positioned (Gerstein & Chothia, 1996). Alternatively, one can make up the positions of missing solvent molecules. These can be placed either according to a regular grid-like arrangement or, more realistically, according to the results of molecular simulation (Finney *et al.*, 1980; Gerstein *et al.*, 1995; Richards, 1974).

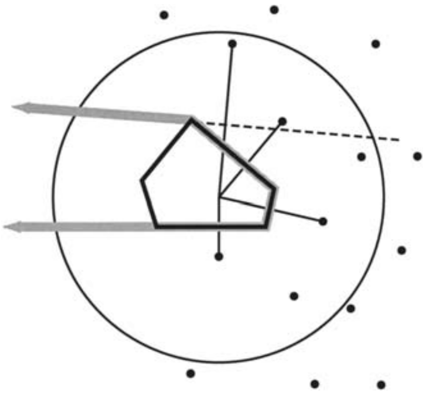


Figure 22.1.3.1
The problem of the protein surface. This figure shows the difficulty in constructing Voronoi polyhedra for atoms on the protein surface. If all the water molecules near the surface are not resolved in a crystal structure, one often does not have enough neighbours to define a closed polyhedron. This figure should be compared with Fig. 22.1.2.1, illustrating the basic Voronoi construction. The two figures are the same except that in this figure, some of the atoms on the left are missing, giving the central atom an open polyhedron. The broken lines indicate planes that were initially included in the polyhedron but then removed by the 'chopping-down' procedure (see Fig. 22.1.2.4).

22.1.3.2. Definitions of surface in terms of Voronoi polyhedra (the convex hull)

More fundamentally, however, the 'problem of the protein surface' indicates how closely linked the definitions of surface and volume are and how the definition of one, in a sense, defines the other. That is, the two-dimensional (2D) surface of an object can be defined as the boundary between two 3D volumes. More specifically, the polyhedral faces defining the Voronoi volume of a collection of atoms also define their surface. The surface of a protein consists of the union of (connected) polyhedra faces. Each face in this surface is shared by one solvent atom and one protein atom (Fig. 22.1.3.2).

Another somewhat related definition is the convex hull, the smallest convex polyhedron that encloses all the atom centres (Fig. 22.1.3.2). This is important in computer-graphics applications and as an intermediary in many geometric constructions related to proteins (Connolly, 1991; O'Rourke, 1994). The convex hull is a subset of the Delaunay triangulation of the surface atoms. It is quickly located by the following procedure (Connolly, 1991): Find the atom farthest from the molecular centre. Then choose two of its neighbours (as determined by the Delaunay triangulation) such that a plane through these three atoms has all the remaining atoms of the molecule on one side of it (the 'plane test'). This is the first triangle in the convex hull. Then one can choose a fourth atom connected to at least two of the three in the triangle and repeat the plane test, and by iteratively repeating this procedure, one can 'sweep' across the surface of the molecule and define the whole convex hull.

Other parts of the Delaunay triangulation can define additional surfaces. The part of the triangulation connecting the first layer of water molecules defines a surface, as does the part joining the second layer. The second layer of water molecules, in fact, has been suggested on physical grounds to be the natural boundary for a protein in solution (Gerstein & Lynden-Bell, 1993*c*). Protein surfaces defined in terms of the convex hull or water layers tend to be 'smoother' than those based on Voronoi faces, omitting deep grooves and clefts (see Fig. 22.1.3.2).

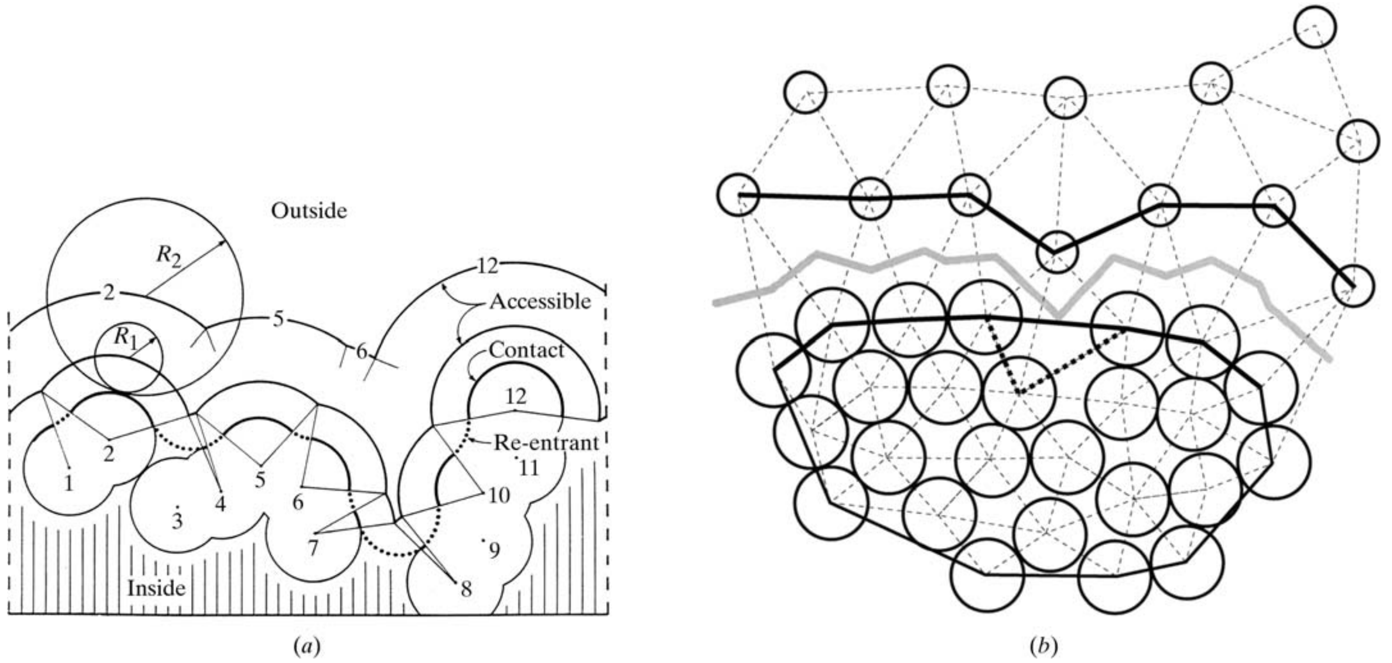


Figure 22.1.3.2
Definitions of the protein surface. (*a*) The classic definitions of protein surface in terms of the probe sphere, the accessible surface and the molecular surface. (This figure is adapted from Richards, 1977). (*b*) Voronoi polyhedra and Delaunay triangulation can also be used to define a protein surface. In this schematic, the large spheres represent closely packed protein atoms and the smaller spheres represent the small loosely packed water molecules. The Delaunay triangulation is shown by dotted lines. Some parts of the triangulation can be used to define surfaces. The outermost part of the triangulation of *just* the protein atoms forms the convex hull. This is indicated by the thick line around the protein atoms. For the convex-hull construction, one imagines that the water is not present. This is highlighted by the thick dotted line, which shows how Delaunay triangulation of the surface atoms in the presence of the water diverges from the convex hull near a deep cleft. Another part of the triangulation, also indicated by thick black lines, connects the first layer of water molecules (those that touch protein atoms). A time-averaged version of this line approximates the accessible surface. Finally, the light thick lines show the Voronoi faces separating the protein surface atoms from the first layer of water molecules. Note how this corresponds approximately to the molecular surface (considering the water positions to be time-averaged). These correspondences between the accessible and molecular surfaces and time-averaged parts of the Voronoi construction are understandable in terms of which part of the probe sphere (centre or point of tangency) is used for the surface definition. The accessible surface is based on the position of the centre of the probe sphere while the molecular surface is based on the points of tangency between the probe sphere and the protein atoms, and these tangent points are similarly positioned to Voronoi faces, which bisect interatomic vectors between solvent and protein atoms.

22.1.3.3. Definitions of surface in terms of a probe sphere

In the absence of solvent molecules to define Voronoi polyhedra, one can define the protein surface in terms of the position of a hypothetical solvent, often called the probe sphere, that 'rolls' around the surface (Richards, 1977) (Fig. 22.1.3.2). The surface of the probe is imagined to be maintained at a tangent to the van der Waals surface of the model.

Various algorithms are used to cause the probe to visit all possible points of contact with the model. The locus of either the centre of the probe or the tangent point to the model is recorded. Either through exact analytical functions or numerical approximations of adjustable accuracy, the algorithms provide an estimate of the area of the resulting surface. (See Chapter 22.2 for a more extensive discussion of the definition, calculation and use of areas.)

Depending on the probe size and whether its centre or point of tangency is used to define the surface, one arrives at a number of commonly used definitions, summarized in Table 22.1.4.1 and Fig. 22.1.3.2.

22.1.3.3.1. van der Waals surface (VDWS)

The area of the van der Waals surface will be calculated by the various area algorithms (see Section 22.2.2 in Chapter 22.2) when the probe radius is set to zero. This is a mathematical calculation only. There is no physical procedure that will measure van der Waals surface area directly. From a mathematical point of view, it is just the first of a set of solvent-accessible surfaces calculated with differing probe radii.

22.1.3.3.2. Solvent-accessible surface (SAS)

The solvent-accessible surface is convex and closed, with defined areas assignable to each individual atom (Lee & Richards, 1971). However, the individual calculated values vary in a complex fashion with variations in the radii of the probe and protein atoms. This radius is frequently, but not always, set at a value considered to represent a water molecule (1.4 Å). The total SAS area increases without bound as the size of the probe increases.

22.1.3.3.3. Molecular surface as the sum of the contact and re-entrant surfaces (MS = CS + RS)

Like the solvent-accessible surface, the molecular surface is also closed, but it contains a mixture of convex and concave patches, the sum of the contact and re-entrant surfaces. The ratio of these two surfaces varies with probe radius. In the limit of infinite probe radius, the molecular surface becomes convex and attains a limiting minimum value (*i.e.* it becomes a convex hull, similar to the one described above). The molecular surface cannot be divided up and assigned unambiguously to individual atoms.

The contact surface is not closed. Instead, it is a series of convex patches on individual atoms, simply related to the solvent-

Table 22.1.4.1
Standard atomic radii (Å)

For '*' see following notes on specific sets of values. *Bondi*: Values assigned on the basis of observed packing in condensed phases (Bondi, 1968). *Lee & Richards*: Values adapted from Bondi (1964) and used in Lee & Richards (1971). *Shrake & Rupley*: Values taken from Pauling (1960) and used in Shrake & Rupley (1973). >C= value can be either 1.5 or 1.85. *Richards*: Minor modification of the original Bondi set in Richards (1974). (Rationale not given.) See original paper for discussion of aromatic carbon value. *Chothia*: From packing in amino-acid crystal structures. Used in Chothia (1975). *Richmond & Richards*: No rationale given for values used in Richmond & Richards (1978). *Gelin & Karplus*: Origin of values not specified. Used in Gelin & Karplus (1979). *Dunfield et al.*: Detailed description of deconvolution of molecular crystal energies. Values represent one-half of the heavy-atom separation at the minimum of the Lennard–Jones 6–12 potential functions for symmetrical interactions. Used in Nemethy *et al.* (1983) and Dunfield *et al.* (1979). *ENCAD*: A set of radii, derived in Gerstein *et al.* (1995), based solely on the *ENCAD* molecular dynamics potential function in Levitt *et al.* (1995). To determine these radii, the separation at which the 6–12 Lennard–Jones interaction energy between equivalent atoms was 0.25 k_BT was determined (0.15 kcal mol^{-1}; 1 kcal = 4.184 kJ). *CHARMM*: Determined in the same way as the *ENCAD* set, but for the *CHARMM* potential (Brooks *et al.*, 1983) (parameter set 19). *Tsai et al.*: Values derived from a new analysis (Tsai *et al.*, 1999) of the most common distances of approach of atoms in the Cambridge Structural Database.

Atom type and symbol		Bondi (1968)	Lee & Richards (1971)	Shrake & Rupley (1973)	Richards (1974)	Chothia (1975)	Richmond & Richards (1978)	Gelin & Karplus (1979)	Dunfield *et al.* (1979)	*ENCAD* derived (1995)	*CHARMM* derived (1995)	Tsai *et al.* (1999)
$-CH_3$	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.95	2.13	1.82	1.88	1.88
$-CH_2-$	Aliphatic, methyl	2.00	1.80	2.00	2.00	1.87	1.90	1.90	2.23	1.82	1.88	1.88
$>CH-$	Aliphatic, CH	—	1.70	2.00	2.00	1.87	1.90	1.85	2.38	1.82	1.88	1.88
$\geq CH=$	Aromatic, CH	—	1.80	1.85	*	1.76	1.70	1.90	2.10	1.74	1.80	1.76
$>C=$	Trigonal, aromatic	1.74	1.80	*	1.70	1.76	1.70	1.80	1.85	1.74	1.80	1.61
$-NH_3^+$	Amino, protonated	—	1.80	1.50	2.00	1.50	0.70	1.75	—	1.68	1.40	1.64
$-NH_2$	Amino or amide	1.75	1.80	1.50	—	1.65	1.70	1.70	—	1.68	1.40	1.64
$>NH$	Peptide, NH or N	1.65	1.52	1.40	1.70	1.65	1.70	1.65	1.75	1.68	1.40	1.64
$=O$	Carbonyl oxygen	1.50	1.80	1.40	1.40	1.40	1.40	1.60	1.56	1.34	1.38	1.42
$-OH$	Alcoholic hydroxyl	—	1.80	1.40	1.60	1.40	1.40	1.70	—	1.54	1.53	1.46
$-OM$	Carboxyl oxygen	—	1.80	1.89	1.50	1.40	1.40	1.60	1.62	1.34	1.41	1.42
$-SH$	Sulfhydryl	—	1.80	1.85	—	1.85	1.80	1.90	—	1.82	1.56	1.77
$-S-$	Thioether or –S–S–	1.80	—	—	1.80	1.85	1.80	1.90	2.08	1.82	1.56	1.77

accessible surface of the same atoms. In complementary fashion, the re-entrant surface is also not closed but is a series of concave patches that is part of the probe surface where it contacts two or three atoms simultaneously. At infinite probe radius, the re-entrant areas are plane surfaces, at which point the molecular surface becomes a convex surface. The re-entrant surface cannot be divided up and assigned unambiguously to individual atoms. Note that the molecular surface is simply the union of the contact and re-entrant surfaces, so in terms of area MS = CS + RS.

22.1.3.3.4. *Further points*

The detail provided by these surfaces will depend on the radius of the probe used for their construction.

One may argue that the behaviour of the rolling probe sphere does not accurately model real hydrogen-bonded water. Instead, its 'rolling' more closely mimics the behaviour of a nonpolar solvent. An attempt has been made to incorporate more realistic hydrogen-bonding behaviour into the probe sphere, allowing for the definition of a hydration surface more closely linked to the behaviour of real water (Gerstein & Lynden-Bell, 1993*c*).

The definitions of accessible surface and molecular surface can be related back to the Voronoi construction. The molecular surface is similar to 'time-averaging' the surface formed from the faces of Voronoi polyhedra (the Voronoi surface) over many water configurations, and the accessible surface is similar to averaging the Delaunay triangulation of the first layer of water molecules over many configurations.

There are a number of other definitions of protein surfaces that are unrelated to either the probe-sphere method or Voronoi polyhedra and provide complementary information (Kuhn *et al.*, 1992; Leicester *et al.*, 1988; Pattabiraman *et al.*, 1995).

22.1.4. Definitions of atomic radii

The definition of protein surfaces and volumes depends greatly on the values chosen for various parameters of linear dimension – in particular, van der Waals and probe-sphere radii.

22.1.4.1. *van der Waals radii*

For all the calculations outlined above, the hard-sphere approximation is used for the atoms. (One must remember that in reality atoms are neither hard nor spherical, but this approximation has a long history of demonstrated utility.) There are many lists of the radii of such spheres prepared by different laboratories, both for single atoms and for unified atoms, where the radii are adjusted to approximate the joint size of the heavy atom and its bonded hydrogen atoms (clearly not an actual spherical unit).

Some of these lists are reproduced in Table 22.1.4.1. They are derived from a variety of approaches, *e.g.* looking for the distances of closest approach between atoms (the Bondi set) and energy calculations (the *CHARMM* set). The differences between the sets often come down to how one decides to truncate the Lennard–Jones potential function. Further differences arise

Table 22.1.4.2
Probe radii and their relation to surface definition

The values of 1.4 and, especially, 10 Å are only approximate. One could, of course, use 1.5 Å for a water radius or 15 Å for a ligand radius, depending on the specific application.

Probe radius (Å)	Part of probe sphere	Type of surface
0	Centre (or tangent)	van der Waals surface (VDWS)
1.4	Centre	Solvent-accessible surface (SAS)
1.4	Tangent (one atom)	Contact surface (CS, from parts of atoms)
1.4	Tangent (two or three atoms)	Re-entrant surface (RS, from parts of probe)
1.4	Tangent (one, two, or three atoms)	Molecular surface (MS = CS + RS)
10	Centre	A ligand- or reagent-accessible surface
∞	Tangent	Minimum limit of MS (related to convex hull)
∞	Centre	Undefined

from the parameterization of water and other hydrogen-bonding molecules, as these substances really should be represented with two radii, one for their hydrogen-bonding interactions and one for their VDW interactions.

Perhaps because of the complexities in defining VDW parameters, there are some great differences in Table 22.1.4.1. For instance, the radius for an aliphatic CH (>CH=) ranges from 1.7 to 2.38 Å, and the radius for carboxyl oxygen ranges from 1.34 to 1.89 Å. Both of these represent at least a 40% variation. Moreover, such differences are practically quite significant, since many geometrical and energetic calculations are very sensitive to the choice of VDW parameters, particularly the relative values within a single list. (Repulsive core interactions, in fact, vary almost exponentially.) Consequently, proper volume and surface comparisons can only be based on numbers derived through use of the same list of radii.

In the last column of the table we give a recent set of VDW radii that has been carefully optimized for use in volume and packing calculations. It is derived from analysis of the most common distances between atoms in small-molecule crystal structures in the Cambridge Structural Database (Rowland & Taylor, 1996; Tsai *et al.*, 1999).

22.1.4.2. The probe radius

A series of surfaces can be described by using a probe sphere with a specified radius. Since this is to be a convenient mathematical construct in calculation, any numerical value may be chosen with no necessary relation to physical reality. Some commonly used examples are listed in Table 22.1.4.2.

The solvent-accessible surface is intended to be a close approximation to what a water molecule as a probe might 'see' (Lee & Richards, 1971). However, there is no uniform agreement on what the proper water radius should be. Usually it is chosen to be about 1.4 Å.

22.1.5. Application of geometry calculations: the measurement of packing

22.1.5.1. Using volume to measure packing efficiency

Volume calculations are principally applied in measuring packing. This is because the packing efficiency of a given atom is simply the ratio of the space it could minimally occupy to the space that it actually does occupy. As shown in Fig. 22.1.5.1, this ratio can be expressed as the VDW volume of an atom divided by its Voronoi volume (Richards, 1974, 1985; Richards & Lim, 1994). (Packing efficiency also sometimes goes by the equivalent terms 'packing density' or 'packing coefficient'.) This simple definition masks considerable complexities – in particular, how does one determine the volume of the VDW envelope (Petitjean, 1994)? This requires knowledge of what the VDW radii of atoms are, a

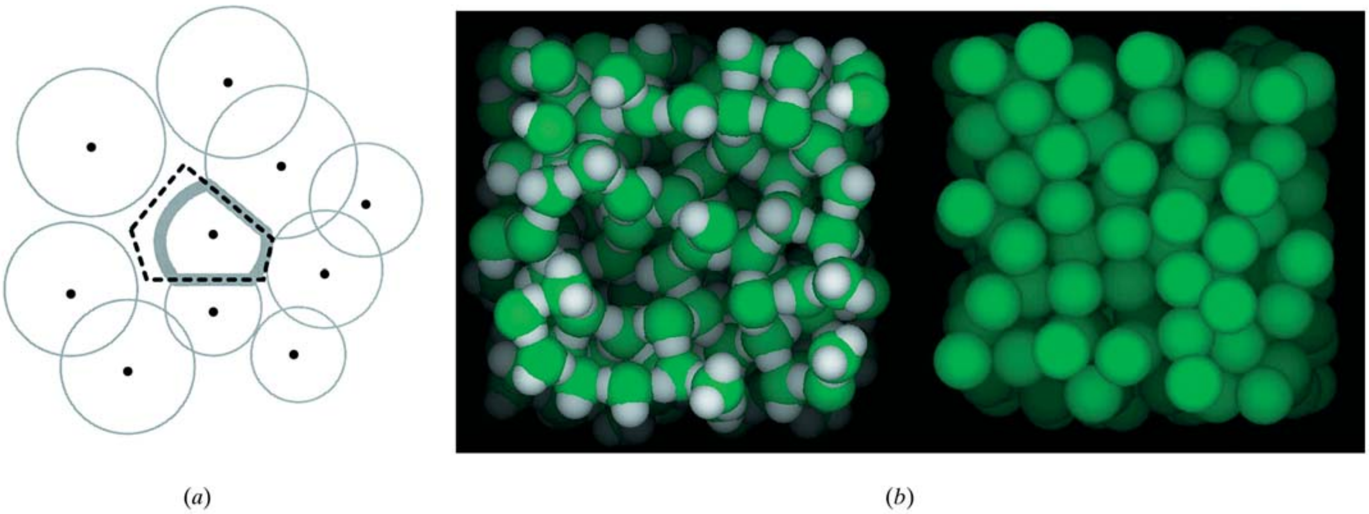


Figure 22.1.5.1
Packing efficiency. (*a*) The relationship between Voronoi polyhedra and packing efficiency. Packing efficiency is defined as the volume of an object as a fraction of the space that it occupies. (It is also known as the 'packing coefficient' or 'packing density'.) In the context of molecular structure, it is measured by the ratio of the VDW volume ($V_{\rm VDW}$, shown by a light grey line) and Voronoi volume ($V_{\rm Vor}$, shown by a dotted line). This calculation gives absolute packing efficiencies. In practice, one usually measures a relative efficiency, relative to the atom in a reference state: $(V_{\rm VDW}/V_{\rm Vor})/[V_{\rm VDW}/V_{\rm Vor}(\text{ref})]$. Note that in this ratio the unchanging VDW volume of an atom cancels out, leaving one with just a ratio of two Voronoi volumes. Perhaps more usefully, when one is trying to evaluate the packing efficiency P at an interface, one computes $P = p\sum V_i/\sum v_i$, where p is packing efficiency of the reference data set (usually 0.74), V_i is the actual measured volume of each atom i at the interface and v_i is the reference volume corresponding to the type of atom i. (*b*) A graphical illustration of the difference between tight packing and loose packing. Frames from a simulation are shown for liquid water (left) and for liquid argon, a simple liquid (right). Owing to its hydrogen bonds, water is much less tightly packed than argon (packing efficiency of 0.35 *versus* ~0.7). Each water molecule has only four to five nearest neighbours while each argon atom has about ten.

Table 22.1.5.1
Standard residue volumes

The mean standard volume, the standard deviation about the mean and the frequency of occurrence of each residue in the protein core are given. Considering cysteine (Cyh, reduced) to be chemically different from cystine (Cys, involved in a disulfide and hence oxidized) gives 21 different residues. These residue volumes are adapted from the *ProtOr* parameter set (also known as the BL+ set) in Tsai *et al.* (1999) and Tsai *et al.* (2001). For this set, the averaging is done over 87 representative high-resolution crystal structures, only buried atoms not in contact with ligands are selected, the radii set shown in the last column of Table 22.1.4.1 is used and the volumes are computed in the presence of the crystal water. The frequencies for buried residues are from Harpaz *et al.* (1994).

Residue	Volume (Å^3)	Standard deviation (Å^3)	Frequency (%)
Ala	89.3	3.5	13
Val	138.2	4.8	13
Leu	163.1	5.8	12
Gly	63.8	2.7	11
Ile	163.0	5.3	9
Phe	190.8	4.8	6
Ser	93.5	3.9	6
Thr	119.6	4.2	5
Tyr	194.6	4.9	3
Asp	114.4	3.9	3
Cys	102.5	3.5	3
Pro	121.3	3.7	3
Met	165.8	5.4	2
Trp	226.4	5.3	2
Gln	146.9	4.3	2
His	157.5	4.3	2
Asn	122.4	4.6	1
Glu	138.8	4.3	1
Cyh	112.8	5.5	1
Arg	190.3	4.7	1
Lys	165.1	6.9	1

subject on which there is not universal agreement (see above), especially for water molecules and polar atoms (Gerstein *et al.*, 1995; Madan & Lee, 1994).

Knowing that the absolute packing efficiency of an atom is a certain value is most useful in a comparative sense, *i.e.* when comparing equivalent atoms in different parts of a protein structure. In taking a ratio of two packing efficiencies, the VDW envelope volume remains the same and cancels. One is left with just the ratio of space that an atom occupies in one environment to what it occupies in another. Thus, for the measurement of packing, standard reference volumes are particularly useful. Recently calculated values of these standard volumes are shown in Tables 22.1.5.1 and 22.1.5.2 for atoms and residues (Tsai *et al.*, 1999).

In analysing molecular systems, one usually finds that close packing is the default (Chandler *et al.*, 1983), *i.e.* atoms pack like billiard balls. Unless there are highly directional interactions (such as hydrogen bonds) that have to be satisfied, one usually achieves close packing to optimize the attractive tail of the VDW interaction. Close-packed spheres of the same size have a packing efficiency of ~0.74. Close-packed spheres of different size are expected to have a somewhat higher packing efficiency. In contrast, water is not close-packed because it has to satisfy the additional constraints of hydrogen bonding. It has an open, tetrahedral structure with a packing efficiency of ~0.35. (This difference in packing efficiency is illustrated in Fig. 22.1.5.1*b*).

22.1.5.2. *The tight packing of the protein core*

The protein core is usually considered to be the atoms inaccessible to solvent *i.e.* with an accessible surface area of zero or a very small number, such as 0.1 Å^2. Packing calculations on the protein core are usually done by calculating the average volumes of the buried atoms and residues in a database of crystal structures. These calculations were first done more than two decades ago (Chothia & Janin, 1975; Finney, 1975; Richards, 1974). The initial calculations revealed some important facts about protein structure. Atoms and residues of a given type inside proteins have a roughly constant (or invariant) volume. This is because the atoms inside proteins are packed together fairly tightly, with the protein interior better resembling a close-packed solid than a liquid or gas. In fact, the packing efficiency of atoms inside proteins is roughly as expected for the close packing of hard spheres (0.74).

More recent calculations measuring the packing in proteins (Harpaz *et al.*, 1994; Tsai *et al.*, 1999) have shown that the packing inside proteins is somewhat tighter (by ~4%) than that observed

Table 22.1.5.2
Standard atomic volumes

Tsai *et al.* (1999) and Tsai *et al.* (2001) clustered all the atoms in proteins into the 18 basic types shown below. Most of these have a simple chemical definition, *e.g.* '=O' are carbonyl carbons. However, some of the basic chemical types, such as the aromatic CH group ('≥CH'), need to be split into two subclusters (bigger and smaller), as is indicated by the column labelled 'Cluster'. Volume statistics were accumulated for each of the 18 types based on averaging over 87 high-resolution crystal structures (in the same fashion as described for the residue volumes in Table 22.1.5.1). No. is the number of atoms averaged over. The final column ('Symbol') gives the standardized symbol used to describe the atom in Tsai *et al.* (1999). The atom volumes shown here are part of the *ProtOr* parameter set (also known as the BL+ set) in Tsai *et al.* (1999).

Atom type	Cluster	Description	Average volume (Å^3)	Standard deviation (Å^3)	No.	Symbol
>C=	Bigger	Trigonal (unbranched), aromatics	9.7	0.7	4184	C3H0b
>C=	Smaller	Trigonal (branched)	8.7	0.6	11876	C3H0s
≥CH	Bigger	Aromatic, CH (facing away from main chain)	21.3	1.9	2063	C3H1b
≥CH	Smaller	Aromatic, CH (facing towards main chain)	20.4	1.7	1742	C3H1s
>CH−	Bigger	Aliphatic, CH (unbranched)	14.4	1.3	3642	C4H1b
>CH−	Smaller	Aliphatic, CH (branched)	13.2	1.0	7028	C4H1s
$-CH_2-$	Bigger	Aliphatic, methyl	24.3	2.1	1065	C4H2b
$-CH_2-$	Smaller	Aliphatic, methyl	23.2	2.3	4228	C4H2s
$-CH_3$		Aliphatic, methyl	36.7	3.2	3497	C4H3u
>N−		Pro N	8.7	0.6	581	N3H0u
>NH	Bigger	Side chain NH	15.7	1.5	446	N3H1b
>NH	Smaller	Peptide	13.6	1.0	10016	N3H1s
$-NH_2$		Amino or amide	22.7	2.1	250	N3H2u
$-NH_3^+$		Amino, protonated	21.4	1.2	8	N4H3u
=O		Carbonyl oxygen	15.9	1.3	7872	O1H0u
−OH		Alcoholic hydroxyl	18.0	1.7	559	O2H1u
−S−		Thioether or –S–S–	29.2	2.6	263	S2H0u
−SH		Sulfhydryl	36.7	4.2	48	S2H1u

initially and that the overall packing efficiency of atoms in the protein core is greater than that in crystals of organic molecules. When molecules are packed this tightly, small changes in packing efficiency are quite significant. In this regime, the limitation on close packing is hard-core repulsion, which is expected to have a twelfth power or exponential dependence, so even a small change is energetically quite substantial. Furthermore, the number of allowable configurations that a collection of atoms can assume without core overlap drops off very quickly as these atoms approach the close-packed limit (Richards & Lim, 1994).

The exceptionally tight packing in the protein core seems to require a precise jigsaw puzzle-like fit of the residues. This appears to be the case for the majority of atoms inside proteins (Connolly, 1986). The tight packing in proteins has, in fact, been proposed as a quality measure in protein crystal structures (Pontius *et al.*, 1996). It is also believed to be a strong constraint on protein flexibility and motions (Gerstein *et al.*, 1993; Gerstein, Lesk & Chothia, 1994). However, there are exceptions, and some studies have focused on these, showing how the packing inside proteins is punctuated by defects, or cavities (Hubbard & Argos, 1994, 1995; Kleywegt & Jones, 1994; Kocher *et al.*, 1996; Rashin *et al.*, 1986; Richards, 1979; Williams *et al.*, 1994). If these defects are large enough, they can contain buried water molecules (Baker & Hubbard, 1984; Matthews *et al.*, 1995; Sreenivasan & Axelsen, 1992).

Surprisingly, despite the intricacies of the observed jigsaw puzzle-like packing in the protein core, it has been shown that one can simply achieve the 'first-order' aspect of this, getting the overall volume of the core right rather easily (Gerstein, Sonnhammer & Chothia, 1994; Kapp *et al.*, 1995; Lim & Ptitsyn, 1970). This has to do with simple statistics for summing random numbers and the fact that the distribution of sizes for amino acids usually found inside proteins is rather narrow (Table 22.1.5.1). In fact, the similarly sized residues Val, Ile, Leu and Ala (with volumes 138, 163, 163 and 89 Å^3) make up about half of the residues buried in the protein core. Furthermore, aliphatic residues, in particular, have a relatively large number of adjustable degrees of freedom per Å^3, allowing them to accommodate a wide range of packing geometries. All of this suggests that many of the features of protein sequences may only require random-like qualities for them to fold (Finkelstein, 1994).

22.1.5.3. Looser packing on the surface

Measuring the packing efficiency inside the protein core provides a good reference point for comparison, and a number of other studies have looked at this in comparison with other parts of the protein. The most obvious thing to compare with the protein inside is the protein outside, or surface. This is particularly interesting from a packing perspective, since the protein surface is covered by water, and water is packed much less tightly than protein and in a distinctly different fashion. (The tetrahedral packing geometry of water molecules gives a packing efficiency of less than half that of hexagonal close-packed solids.)

Calculations based on crystal structures and simulations have shown that the protein surface has intermediate packing, being packed less tightly than the core but not as loosely as liquid water (Gerstein & Chothia, 1996; Gerstein *et al.*, 1995). One can understand the looser packing at the surface than in the core in terms of a simple trade-off between hydrogen bonding and close packing, and this can be explicitly visualized in simulations of the packing in simple toy systems (Gerstein & Lynden-Bell, 1993*a*,*b*).

MG acknowledges support from the NSF Database Activities Program DBI-9723182.

References

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Bernal, J. D. & Finney, J. L. (1967). *Random close-packed hard-sphere model II. Geometry of random packing of hard spheres. Discuss. Faraday Soc.* **43**, 62–69.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–451.

Bondi, A. (1968). *Molecular crystals, liquids and glasses*. New York: Wiley.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem.* **4**, 187–217.

Chandler, D., Weeks, J. D. & Andersen, H. C. (1983). *van der Waals picture of liquids, solids, and phase transformations. Science*, **220**, 787–794.

Chothia, C. (1975). *Structural invariants in protein folding. Nature (London)*, **254**, 304–308.

Chothia, C. & Janin, J. (1975). *Principles of protein–protein recognition. Nature (London)*, **256**, 705–708.

Connolly, M. (1986). *Measurement of protein surface shape by solid angles. J. Mol. Graphics*, **4**, 3–6.

Connolly, M. L. (1991). *Molecular interstitial skeleton. Comput. Chem.* **15**, 37–45.

Dunfield, L. G., Burgess, A. W. & Scheraga, H. A. (1979). *Energy parameters in polypeptides. 8. Empirical potential energy algorithm for the conformational analysis of large molecules. J. Phys. Chem.* **82**, 2609.

Edelsbrunner, H., Facello, M. & Liang, J. (1996). *On the Definition and Construction of Pockets in Macromolecules*, pp. 272–287. Singapore: World Scientific.

Edelsbrunner, H., Facello, M., Ping, F. & Jie, L. (1995). *Measuring proteins and voids in proteins. Proc. 28th Hawaii Intl Conf. Sys. Sci.* pp. 256–264.

Edelsbrunner, H. & Mucke, E. (1994). *Three-dimensional alpha shapes. ACM Trans. Graphics*, **13**, 43–72.

Finkelstein, A. (1994). *Implications of the random characteristics of protein sequences for their three-dimensional structure. Curr. Opin. Struct. Biol.* **4**, 422–428.

Finney, J. L. (1975). *Volume occupation, environment and accessibility in proteins. The problem of the protein surface. J. Mol. Biol.* **96**, 721–732.

Finney, J. L., Gellatly, B. J., Golton, I. C. & Goodfellow, J. (1980). *Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. Biophys. J.* **32**, 17–33.

Gelin, B. R. & Karplus, M. (1979). *Side-chain torsional potentials: effect of dipeptide, protein, and solvent environment. Biochemistry*, **18**, 1256–1268.

Gellatly, B. J. & Finney, J. L. (1982). *Calculation of protein volumes: an alternative to the Voronoi procedure. J. Mol. Biol.* **161**, 305–322.

Gerstein, M. & Chothia, C. (1996). *Packing at the protein–water interface. Proc. Natl Acad. Sci. USA*, **93**, 10167–10172.

Gerstein, M., Lesk, A. M., Baker, E. N., Anderson, B., Norris, G. & Chothia, C. (1993). *Domain closure in lactoferrin: two hinges produce a see-saw motion between alternative close-packed interfaces. J. Mol. Biol.* **234**, 357–372.

Gerstein, M., Lesk, A. M. & Chothia, C. (1994). *Structural mechanisms for domain movements. Biochemistry*, **33**, 6739–6749.

Gerstein, M. & Lynden-Bell, R. M. (1993*a*). *Simulation of water around a model protein helix. 1. Two-dimensional projections of solvent structure. J. Phys. Chem.* **97**, 2982–2991.

Gerstein, M. & Lynden-Bell, R. M. (1993*b*). *Simulation of water around a model protein helix. 2. The relative contributions of packing, hydrophobicity, and hydrogen bonding. J. Phys. Chem.* **97**, 2991–2999.

Gerstein, M. & Lynden-Bell, R. M. (1993*c*). *What is the natural boundary for a protein in solution? J. Mol. Biol.* **230**, 641–650.

Gerstein, M., Sonnhammer, E. & Chothia, C. (1994). *Volume changes on protein evolution. J. Mol. Biol.* **236**, 1067–1078.

Gerstein, M., Tsai, J. & Levitt, M. (1995). *The volume of atoms on the protein surface: calculated from simulation, using Voronoi polyhedra*. *J. Mol. Biol.* **249**, 955–966.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Volume changes on protein folding*. *Structure*, **2**, 641–649.

Hubbard, S. J. & Argos, P. (1994). *Cavities and packing at protein interfaces*. *Protein Sci.* **3**, 2194–2206.

Hubbard, S. J. & Argos, P. (1995). *Evidence on close packing and cavities in proteins*. *Curr. Opin. Biotechnol.* **6**, 375–381.

Kapp, O. H., Moens, L., Vanfleteren, J., Trotman, C. N. A., Suzuki, T. & Vinogradov, S. N. (1995). *Alignment of 700 globin sequences: extent of amino acid substitution and its correlation with variation in volume*. *Protein Sci.* **4**, 2179–2190.

Kleywegt, G. J. & Jones, T. A. (1994). *Detection, delineation, measurement and display of cavities in macromolecular structures*. *Acta Cryst.* D**50**, 178–185.

Kocher, J. P., Prevost, M., Wodak, S. J. & Lee, B. (1996). *Properties of the protein matrix revealed by the free energy of cavity formation*. *Structure*, **4**, 1517–1529.

Kuhn, L. A., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). *The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures*. *J. Mol. Biol.* **228**, 13–22.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility*. *J. Mol. Biol.* **55**, 379–400.

Leicester, S. E., Finney, J. L. & Bywater, R. P. (1988). *Description of molecular surface shape using Fourier descriptors*. *J. Mol. Graphics*, **6**, 104–108.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). *Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution*. *Comput. Phys. Comm.* **91**, 215–231.

Lim, V. I. & Ptitsyn, O. B. (1970). *On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins*. *Mol. Biol. (USSR)*, **4**, 372–382.

Madan, B. & Lee, B. (1994). *Role of hydrogen bonds in hydrophobicity: the free energy of cavity formation in water models with and without the hydrogen bonds*. *Biophys. Chem.* **51**, 279–289.

Matthews, B. W., Morton, A. G. & Dahlquist, F. W. (1995). *Use of NMR to detect water within nonpolar protein cavities*. (Letter.) *Science*, **270**, 1847–1849.

Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1983). *Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions and hydrogen bond interactions for the naturally occurring amino acids*. *J. Phys. Chem.* **87**, 1883–1887.

O'Rourke, J. (1994). *Computational Geometry in C*. Cambridge University Press.

Pattabiraman, N., Ward, K. B. & Fleming, P. J. (1995). *Occluded molecular surface: analysis of protein packing*. *J. Mol. Recognit.* **8**, 334–344.

Pauling, L. (1960). *The Nature of the Chemical Bond*, 3rd ed. Ithaca: Cornell University Press.

Peters, K. P., Fauck, J. & Frommel, C. (1996). *The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria*. *J. Mol. Biol.* **256**, 201–213.

Petitjean, M. (1994). *On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects*. *J. Comput. Chem.* **15**, 1–10.

Pontius, J., Richelle, J. & Wodak, S. J. (1996). *Deviations from standard atomic volumes as a quality measure for protein crystal structures*. *J. Mol. Biol.* **264**, 121–136.

Procacci, P. & Scateni, R. (1992). *A general algorithm for computing Voronoi volumes: application to the hydrated crystal of myoglobin*. *Int. J. Quant. Chem.* **42**, 151–152.

Rashin, A. A., Iofin, M. & Honig, B. (1986). *Internal cavities and buried waters in globular proteins*. *Biochemistry*, **25**, 3619–3625.

Richards, F. M. (1974). *The interpretation of protein structures: total volume, group volume distributions and packing density*. *J. Mol. Biol.* **82**, 1–14.

Richards, F. M. (1977). *Areas, volumes, packing, and protein structure*. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richards, F. M. (1979). *Packing defects, cavities, volume fluctuations, and access to the interior of proteins. Including some general comments on surface area and protein structure*. *Carlsberg Res. Commun.* **44**, 47–63.

Richards, F. M. (1985). *Calculation of molecular volumes and areas for structures of known geometry*. *Methods Enzymol.* **115**, 440–464.

Richards, F. M. & Lim, W. A. (1994). *An analysis of packing in the protein folding problem*. *Q. Rev. Biophys.* **26**, 423–498.

Richmond, T. J. & Richards, F. M. (1978). *Packing of alpha-helices: geometrical constraints and contact areas*. *J. Mol. Biol.* **119**, 537–555.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii*. *J. Phys. Chem.* **100**, 7384–7391.

Shrake, A. & Rupley, J. A. (1973). *Environment and exposure to solvent of protein atoms. Lysozyme and insulin*. *J. Mol. Biol.* **79**, 351–371.

Sibbald, P. R. & Argos, P. (1990). *Weighting aligned protein or nucleic acid sequences to correct for unequal representation*. *J. Mol. Biol.* **216**, 813–818.

Singh, R. K., Tropsha, A. & Vaisman, I. I. (1996). *Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues*. *J. Comput. Biol.* **3**, 213–222.

Sreenivasan, U. & Axelsen, P. H. (1992). *Buried water in homologous serine proteases*. *Biochemistry*, **31**, 12785–12791.

Tsai, J., Gerstein, M. & Levitt, M. (1996). *Keeping the shape but changing the charges: a simulation study of urea and its isosteric analogues*. *J. Chem. Phys.* **104**, 9417–9430.

Tsai, J., Gerstein, M. & Levitt, M. (1997). *Estimating the size of the minimal hydrophobic core*. *Protein Sci.* **6**, 2606–2616.

Tsai, J., Taylor, R., Chothia, C. & Gerstein, M. (1999). *The packing density in proteins: standard radii and volumes*. *J. Mol. Biol.* **290**, 253–266.

Tsai, J., Voss, N. & Gerstein, M. (2001). *Voronoi calculations of protein volumes: sensitivity analysis and parameter database*. *Bioinformatics*. In the press.

Voronoi, G. F. (1908). *Nouvelles applications des paramétres continus à la théorie des formes quadratiques*. *J. Reine Angew. Math.* **134**, 198–287.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins*. *Protein Sci.* **3**, 1224–1235.

International Tables for Crystallography (2012). Vol. F, Chapter 22.2, pp. 713–720.

Chapter 22.2. Molecular surfaces: calculations, uses and representations

M. S. Chapman and M. L. Connolly

22.2.1. Introduction

22.2.1.1. Uses of surface-area calculations

Interactions between molecules are most likely to be mediated by the properties of residues at their surfaces. Surfaces have figured prominently in functional interpretations of macromolecular structure. Which residues are most likely to interact with other molecules? What are their properties: charged, polar, or hydrophobic? What would be the estimated energy of interaction? How do the shapes and properties complement one another? Which surfaces are most conserved among a homologous family? At the centre of these questions that are often asked at the start of a structural interpretation lies the calculation of the molecular and/or accessible surfaces.

Surface-area calculations are used in two ways. Graphical surface representations help to obtain a quick intuitive understanding of potential molecular functions and interactions through visualization of the shape, charge distribution, polarity, or sequence conservation on the molecular surface (for example). Quantitative calculations of surface area are used *en route* to approximations of the free energy of interactions in binding complexes.

Part of this subject area was the topic of an excellent review by Richards (1985), to which the reader is referred for greater coverage of many of the methods of calculation. This review will attempt to incorporate more recent developments, particularly in the use of graphics, both realistic and schematic.

22.2.1.2. Molecular, solvent-accessible and occluded surface areas

The concept of molecular surface derives from the behaviour of non-bonded atoms as they approach each other. As indicated by the Lennard–Jones potential, strong unfavourable interactions of overlapping non-bonding electron orbitals increase sharply according to $1/r^{12}$, and atoms behave almost as if they were hard spheres with *van der Waals* radii that are characteristic for each atom type and nearly independent of chemical context. Of course, when orbitals combine in a covalent bond, atoms approach much more closely. Lower-energy attractions between atoms, such as hydrogen bonds or aromatic ring stacking, lead to modest reductions in the distance of closest approach. The van der Waals surface is the area of a volume formed by placing van der Waals spheres at the centre of each atom in a molecule.

Non-bonded atoms of the same molecule contact each other over (at most) a very small proportion of their van der Waals surface. The surface is complicated with gaps and crevices. Much of this surface is inaccessible to other atoms or molecules, because there is insufficient space to place an atom without resulting in forbidden overlap of non-bonded van der Waals spheres (Fig. 22.2.1.1). These crevices are excluded in the *molecular surface area*. The molecular surface area, also known as the solvent-excluding surface, is the outer surface of the volume from which solvent molecules are excluded. Strictly, this would depend on the orientation of non-spherically symmetric solvents such as water. However, since hydrogen atoms are smaller than oxygen atoms, for current purposes it is sufficient to consider water as a sphere with a radius of 1.4 to 1.7 Å, approximating the 'average' distance from the centre of the oxygen atom to the van der Waals surface of water. The practical definition of the molecular surface is, then, the area of the volume excluded to a spherical probe of 1.4 to 1.7 Å radius.

As an aside, it is important to note that surface-area calculations depend on inexact parameterization. For example, there is no radius of any hard-sphere model that can give a realistic representation of the solvent. Furthermore, the choice of van der Waals radii can depend on whether the distance of zero or minimum potential energy is estimated and the potential-energy function or experimental data used. (Tables of common values are given by Gerstein & Richards in Chapter 22.1.) Thus, calculations of molecular and accessible surfaces are approximate. However, when the errors are averaged over large areas of a macromolecule, the numbers can be precise enough to give important insights into function.

Fig. 22.2.1.1 shows that the molecular surface consists of two components. The contact surface is part of the van der Waals surface. The re-entrant surface encloses the interstitial volume and has components that are the exterior surfaces of atoms (contact surface) and parts of the surfaces of probes placed in positions where they are in contact with van der Waals surfaces of two or more atoms (re-entrant surface).

The occluded molecular surface is an approximate complement to the solvent-accessible surface. It is the part of the surface that would be inaccessible to solvent because of steric conflict with neighbouring macromolecular atoms. It is an approximation in that current calculations use van der Waals surfaces, ignoring the differences between atomic and re-entrant surfaces (see below), and the volume of the probe is not fully accounted for (Pattabiraman *et al.*, 1995). Occluded area is defined as the atomic area whose normals cannot be extended 2.8 Å (the presumptive diameter of a water molecule) without intersecting the van der Waals volume of another atom. This crude approximation to the surface that is inaccessible to water not only

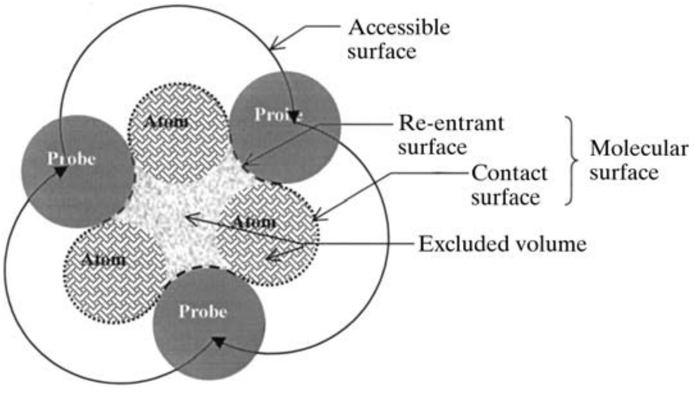


Figure 22.2.1.1
Surfaces in a plane cut through a hypothetical molecule. The molecular surface consists of the sum of the atomic surfaces that can be contacted by solvent molecules and the surface of the space between atoms from which solvent molecules are excluded. The solvent-accessible surface is the surface formed by the set of the centres of spheres that are in closest contact with the molecular surface.

Copyright © 2012 International Union of Crystallography

increases the speed of calculation, but enables surface areas to be partitioned between the atoms. It is used primarily to evaluate model protein structures by comparing the fraction of each amino acid's surface area that is occluded with that calculated for the same residue types in a database of accurate structures.

22.2.1.3. Hydration surface

Whether graphically displaying a molecule or examining potential docking interactions, it is usually the molecular surface or solvent-accessible surface that is used. However, macromolecules also interact through the small (solvent) molecules that are more or less tightly bound (Gerstein & Lynden-Bell, 1993). There is a gradation of how tightly solvent molecules are bound and how many are bound around different side chains. With dynamics simulations, Gerstein & Lynden-Bell (1993) showed that the second hydration shell was a reasonable, practical 'average' limit to which water atoms should be considered significantly perturbed by the protein. They defined a hydration surface as the surface of this second shell and presented evidence that it approximates the boundary between bound and bulk solvent. They presented calculations that showed that molecules interact significantly when their hydration surfaces interact, and not just when they are close enough for their molecular surfaces to form contacts. It may be computationally impractical to perform the simulations required to calculate the hydration surfaces of many proteins, but this work reminds us that energetically significant interactions occur over a wider area than the commonly computed contact molecular-surface area.

22.2.1.4. Hydrophobicity

The hydrophobic effect (Kauzmann, 1959; Tanford, 1997) has its origins in unfavourable entropic terms for water molecules immediately surrounding a hydrophobic group. In the bulk solvent, each water molecule can be oriented in a variety of ways with favourable hydrogen bonding. At the interface with a hydrophobic group, hydrogen bonds are possible only in some directions, with some configurations of the water molecules. When a hydrophobic group is embedded in water, the surrounding solvent molecules have a more restricted set of hydrogen-bonding configurations, resulting in an unfavourable entropic term. The magnitude of the entropic term should be proportional to the number of solvent molecules immediately surrounding the hydrophobic group. This integer number can be considered very approximately proportional to the area of the surface made by the centres of the set of possible solvent probes contacting the solute, *i.e.* the *solvent-accessible surface area* (Fig. 22.2.1.1). When large areas are considered, summed over many hydrophobic atoms, the errors of this non-integer approximation are insignificant. It is now common practice to estimate the hydrophobic effect free-energy contribution by multiplying the change in macromolecular surface area by an energy per unit area [~80 J mol^{-1} Å^{-2} (Richards, 1985), but see also below].

22.2.2. Calculation of surface area and energies of interaction

22.2.2.1. Introduction

The first method to be discussed allows the calculation of an accessible surface. The first method for calculating molecular surface involved raining water down on a model of a macromolecule and constructing a surface by making a net under the spheres in their landing positions (Greer & Bush, 1978). This ignored overhangs and was replaced by the dot surface method. More recently, methods were developed to make polyhedral surfaces of triangles by contouring between lattice points or by delimiting with arcs the spherical and toroidal surfaces and then subdividing the piece-wise quartic molecular surface. The surface is then composed of patches whose areas can be precisely integrated. van der Waals surfaces consist of convex spherical triangles whose areas can be estimated by the Gauss–Bonnet theorem. Re-entrant surfaces are comprised of concave spherical triangles whose areas can be similarly estimated and toroidal saddle-shaped patches whose areas can be calculated by analytical geometry and calculus.

22.2.2.2. Lee & Richards planar slices

The first method for calculating the accessible surface area overlaid the molecule on a regular stack of finely spaced parallel planes (Lee & Richards, 1971). The advantage of this method was the ease with which the area could be calculated. The intersection of the atomic surfaces with the planes were circular arcs whose lengths were readily calculated and multiplied by the planar spacing to give an approximation to the surface area. Programs that are currently distributed use more sophisticated methods.

22.2.2.3. Connolly dot surface algorithm

A molecular dot surface is a smooth envelope of points on the molecular surface. A probe sphere is placed at a set of approximately evenly spaced points so that the probe and van der Waals surfaces of a given atom are tangential. If the probe sphere does not overlap any other atom, the point is designated as surface. To define the re-entrant surface, sphere centres are also sampled that are tangential to both van der Waals spheres of a pair of neighbouring atoms and are equidistant from the interatomic axis. Arcs are then drawn between surface points and the arcs are subdivided into a set of finely spaced points to define the re-entrant surface. Similarly, spheres contacting triplets of neighbouring atoms are tested, and approximately evenly spaced points within the concave triangle defined by the three contact points are added to the re-entrant surface.

22.2.2.4. Marching-cube algorithm

This is conceptually the simplest method and is used in the program *GRASP* (Nicholls *et al.*, 1991). First, grid points of a cubic lattice overlaid on the molecule are segregated into 'interior' and 'exterior' as follows. All points farther from an atom than the sum of the van der Waals radius and a probe radius are flagged as external. External points with an internal neighbour are flagged as an approximate 'accessible surface'. All grid points falling within probe spheres centred at each surface point now join the set of exterior points. Points that remain 'interior' define the volume enclosed by the molecular surface.

All that remains is to contour the molecular surface that lies between interior and exterior grid points. It is a little complicated in three dimensions and is achieved by the marching-cube algorithm. Cubes containing adjacent grid points that are both interior and exterior are used to define potential polyhedral vertices. Triangles are defined by joining the midpoints of unit-cell edges that have one interior and one exterior point. The triangles are joined at their edges in a consistent manner to create a polyhedral surface.

22.2.2.5. Complete and connected rolling algorithms

Several algorithms start by dividing the surface into regions within which the surface is smooth and continuous. The surface can be efficiently described in terms of a set of arcs and their start and end points. In complete rolling, the probe is placed in all possible positions at which it contacts the van der Waals spheres of three neighbouring atoms. Those surrounding the same atom are paired as the start and end points of an arc. The complete rolling algorithm does not distinguish outer and inner (cavity) surfaces. In the connected rolling algorithm, the process starts at a triple contact point that is far from the centre of mass and therefore likely to be external. The probe is then rolled only along crevices between two atoms, pursuing all alternatives, stopping each pathway only when the probe returns to a place that has already been probed. This algorithm therefore produces only the outer surface.

22.2.2.6. Analytic surface calculations and the Gauss–Bonnet theorem

An analytical method was also proposed for calculating approximate accessible areas (Wodak & Janin, 1980). It assumed random distributions of neighbouring atoms, but this can be a sufficient approximation when calculating the area of an entire molecule. The areas of spherical and toroidal pieces of surface can be calculated exactly by analytic and differential geometry (Richmond, 1984; Connolly, 1983). An advantage of analytical expressions over the prior numerical approximations is that analytical derivatives of the areas can be calculated, albeit with significant difficulty. This then provides the opportunity to optimize atomic positions with respect to surface area. Pseudo-energy functions that approximate the hydrophobic contribution to free energy with a term proportional to the accessible surface area (Richards, 1977) can therefore be incorporated in energy-minimization programs. Although rigorous, these methods are computationally cumbersome and are not used in all energy-minimization routines. Incorporation of solvent effects may become more universal with the Gaussian atom approximations discussed below.

22.2.2.7. Approximations to the surface

The methods discussed above are computationally quite cumbersome, especially if they need to be repeated many times. Thus, they are not well suited to comparisons of many structures. They are also not well suited to the calculation of surface-area-dependent energy terms during dynamics simulation or energy minimization, which require the calculation of the derivatives of the surface area with respect to atomic position. It has been argued by several (including A. Nicholls and K. Sharp, personal communications) that simplifying approximations to the surface-area calculations are in order, because the common uses of surface area already embody crude *ad hoc* approximations, such as non-integer numbers of spherical solvent molecules.

In the treatments discussed earlier, the volume of the protein is (implicitly) described by a set of overlapping step functions that have a constant value if close enough to an atom, or zero if not. Several authors have replaced these step functions with continuous spherical Gaussian functions centred on each atom (Gerstein, 1992; Grant & Pickup, 1995) in treatments reminiscent of Ten Eyck's electron-density calculations (Ten Eyck, 1977). This speeds up the calculation and also facilitates the calculation of analytical derivatives of the surface area. A surface can be calculated for graphical display by contouring the continuous function at an appropriate threshold. The final envelope can be modified by using iterative procedures that fill cavities and crevices that are (nearly) surrounded by protein atoms (Gerstein, 1992).

Table 22.2.3.1
The atomic solvation parameters of Eisenberg & McLachlan (1986)

Atom	$\Delta\sigma$(atom) (J mol^{-1} Å^{-2})
C	67 (8)
N/O	−25 (17)
$O^{\cdot\cdot}$	−101 (42)
N^{+}	−210 (38)
S	88 (42)

22.2.2.8. Extended atoms account for missing hydrogen atoms

Structures of macromolecules determined by X-ray crystallography rarely reveal the positions of the hydrogen atoms. It is, of course, possible to add explicit hydrogen atoms at the stereochemically most likely positions, but this is rarely done for surface-area calculations. Instead, their average effect is approximately and implicitly accounted for by increasing the heteroatom van der Waals radius by 0.1 to 0.3 Å. (It is not usual to smear atoms to account for thermal motion.)

22.2.3. Estimation of binding energies

22.2.3.1. Hydrophobicity

As previously introduced, hydrophobic energies result primarily from the increased entropy of water molecules at the macromolecule–solvent interface and can be estimated from the accessible surface area. A number of different constants relating area to free energy of transfer from a hydrophobic to aqueous environment have been proposed in the range of 67 to 130 J mol^{-1} Å^{-2} (Reynolds *et al.*, 1974; Chothia, 1976; Hermann, 1977; Eisenberg & McLachlan, 1986), but if a single value is to be used for all of the protein surface, the consensus among crystallographers has been about 80 J mol^{-1} Å^{-2} (Richards, 1985).

There are two widely used enhancements of the basic method. Atomic solvation parameters (ASPs, $\Delta\sigma$) remove the assumption that all protein atoms have equal potential influence on the hydrophobic free energy. Eisenberg & McLachlan (1986) determined separate ASPs for atom types C, N/O, $O^{\cdot\cdot}$, N^{+} and S (treating hydrogen atoms implicitly) by fitting these constants to the experimentally determined octanol/water relative transfer free energies of the 20 amino-acid side chains of Fauchere & Pliska (1983), assuming standard conformations of the side chains. A much improved free energy change of solvation can then be estimated from $\Delta G = \sum_{\mathrm{atoms}\, i} \Delta\sigma_i A_i$, where the summation is over all atoms with accessible area A and $\Delta\sigma_i$ is specific for the atom type. Their estimates of ASPs are given in Table 22.2.3.1. Use of ASPs rather than a single value for all atoms makes substantial differences to the estimated free energies of association of macromolecular assemblies (Xie & Chapman, 1996). Through calculation of the overall energy of solvation, calculations with ASPs also allow discrimination between proposed structures that are correctly folded (with hydrophobic side chains that are predominantly internal) and those that are not (Eisenberg & McLachlan, 1986).

The work of Sharp *et al.* (1991) indicates that hydrophobicity depends not only on surface area, but curvature. Sharp *et al.* were trying to reconcile long-apparent differences between microscopic and macroscopic measurements of hydrophobicity (Tanford, 1979). Microscopic measurements, the basis of all of our preceding discussions, are derived from the partitioning of dilute solutes between solvents. Macroscopic values can come from the measurements of the surface tension between a liquid bulk of the molecule of interest and water. Macroscopic values for aliphatic carbons are much higher, ~302 J mol^{-1} Å^{-2}. Postulating that the entropic effects at the heart of hydrophobicity depended on the number of water molecules in contact with each other at the molecular surface (Nicholls *et al.*, 1991), Sharp *et al.* pointed out that not all surfaces were equivalent. Relative to a plane, concave solute surfaces would accommodate fewer solvent molecules neighbouring the molecular surface, whereas convex surfaces would accommodate more. Their treatment could be considered to be a second-order approximation to the number of interfacial solvent molecules, compared to the prior first-order consideration of only area.

To calculate the curvature of point *a* on the accessible surface (relative to that of a plane), a sphere of twice the solvent radius is drawn (Nicholls *et al.*, 1991). This represents the locus of the centres of solvent molecules that could be in contact with a solvent at *a*. A curvature correction, *c*, is the proportion of points on the spherical surface that are inside the inaccessible volume, relative to that for a planar accessible surface ($\frac{1}{2}$). In calculating the free energy of transfer, each element of the accessible area is multiplied by its curvature correction. When this is done, the increasingly convex surfaces of small aliphatic molecules account for most of the discrepancy between microscopic and macroscopic hydrophobicities (Nicholls *et al.*, 1991). Furthermore, it emphasizes that, just by their shape, concave surfaces can become relatively hydrophobic. This has been clearly illustrated with *GRASP* surface representations (see below) in which the accessible surface is coloured according to the local curvature (Nicholls *et al.*, 1991). Consideration of curvature also indicates that the energy of macromolecular association is slightly less than it would otherwise be due to the generation of a concave collar at the interface between two binding macromolecules (Nicholls *et al.*, 1991).

22.2.3.2. *Estimates of binding energies*

In a molecular association in which (as is often the case) hydrophobic interactions dominate, the binding energy can be estimated from the surfaces of the individual molecules that become buried upon association (Richards, 1985). The buried area is simply the sum of the surfaces of the two molecules (calculated independently) minus the surface of the complex, calculated as if one molecule. Usually, all heteroatoms are regarded as equivalent, and the buried area is multiplied by a uniform constant, say 80 J mol^{-1} Å^{-2} (Richards, 1985). It is only slightly more complicated to use the different ASPs (Eisenberg & McLachlan, 1986) for different atom types and/or to account for curvature (Nicholls *et al.*, 1991). It should be noted that in many crystal structures, the distinction between atom types in some side chains remains indeterminate, *e.g.* N and C in histidines, O and O$^{\cdot\cdot}$ in carboxylates, and N and N^{+} in arginines. In such cases, average values of the two ASPs can be used (Xie & Chapman, 1996). Such energy calculations have been put to several uses, including attempts to predict assembly and disassembly pathways for viral capsid assemblies (Arnold & Rossmann, 1990; Xie & Chapman, 1996, and citations therein).

22.2.3.3. *Other non-graphical interpretive methods using surface area*

Which are the amino acids most likely to interact with other molecules? It is reasonable to expect them to be surface-accessible. In determining which residues are most surface-exposed, it is necessary to partition molecular or accessible surfaces between atoms. Contact surfaces (Fig. 22.2.1.1) are atom specific. Re-entrant or accessible surfaces can be divided among surface atoms by proximity. Surface areas can then be summed over the atoms in a residue. Accessible surface areas are sometimes reported as *accessibilities* (Lee & Richards, 1971) – fractions of a maximum where the standard is evaluated from a tripeptide in which the residue of interest is surrounded by glycines. A different approach to assessing surface exposure is to ask what is the largest molecular fragment that could contact a given atom. This is commonly assayed by determining the largest sphere that can be placed tangentially to the van der Waals surface without intersecting any other atom. An alternative approach to locating functionally important surface regions was proposed in the mid-1980s, but is currently not used very often. The local irregularity of surface texture was characterized through measurement of the fractal dimension (Lewis & Rees, 1985).

Substrates, drugs and ligands often bind in clefts or pockets that are concave in shape. Conversely, it is the most exposed convex regions that are likely to be antigenic. The surface shape can be determined by placing a large (say 6 Å radius) sphere at each vertex of the polyhedral molecular surface. If more than half of the sphere's volume overlaps the molecular volume, then the surface is concave, while if less than half, the surface is convex.

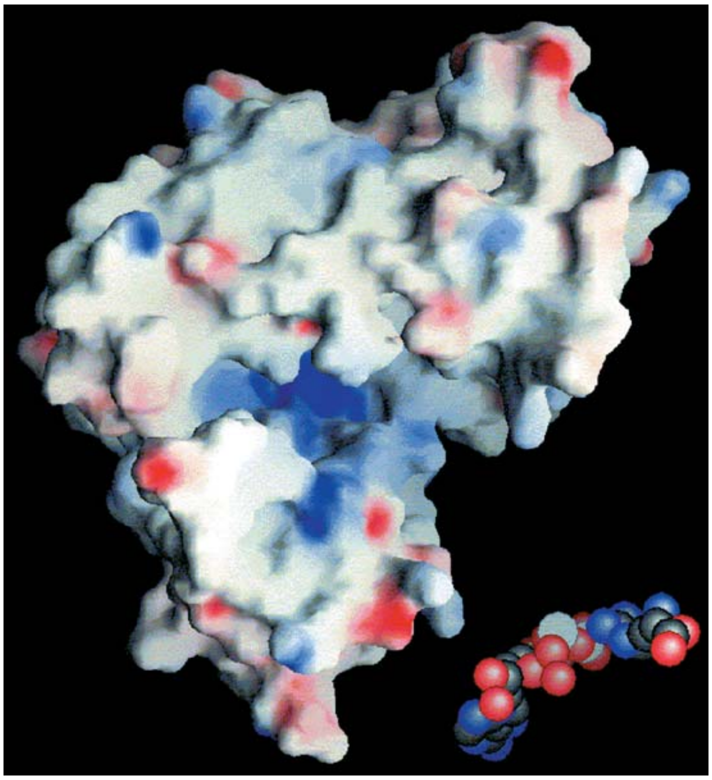
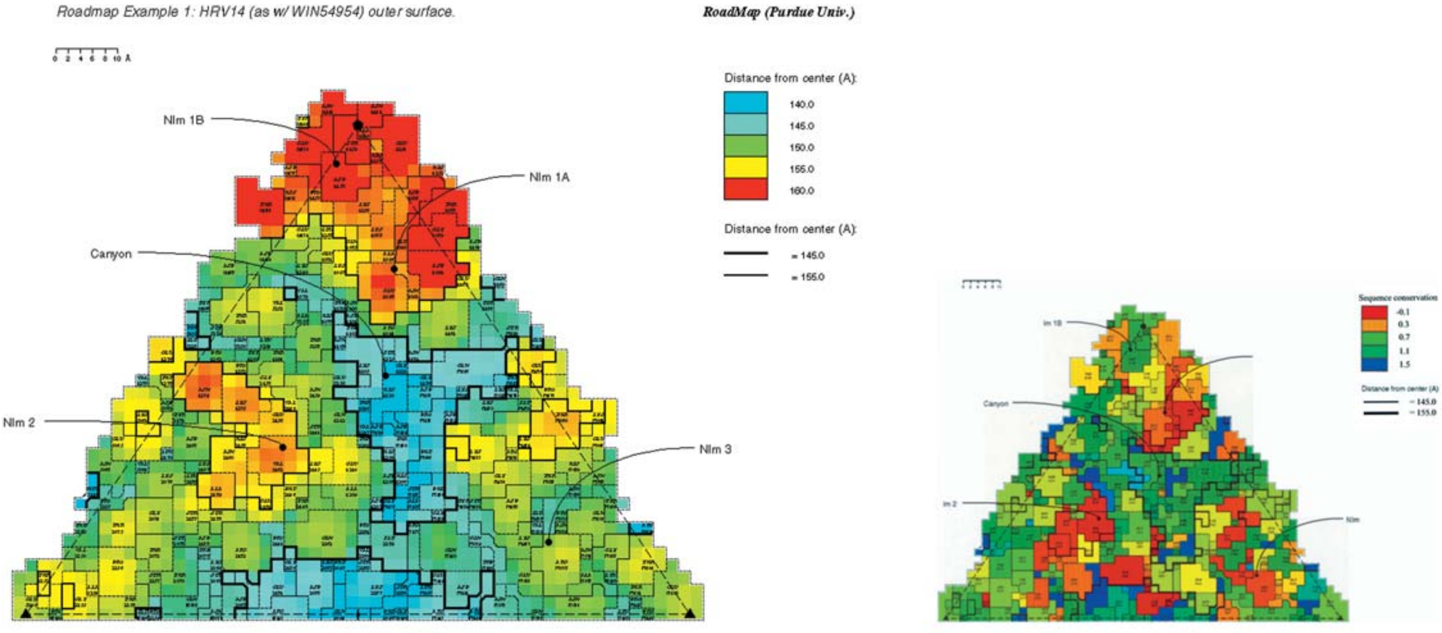


Figure 22.2.4.1
GRASP example. The larger picture shows the molecular surface of arginine kinase (Zhou *et al.*, 1998) with the domains and a loop moved to the open configuration seen in a homologous creatine kinase structure (Fritz-Wolf *et al.*, 1996). The surface, coloured with positive charge blue and negative charge red, demonstrates that the active-site pocket (centre) is the most positively charged part of the structure. It complements the negatively charged phosphates of the transition-state analogue components that are shown, moved as a rigid body to the bottom right. They are shown in van der Waals representation, in which oxygens are red, carbons black and nitrogens blue.



(*a*) (*b*)

Figure 22.2.4.2

(*a*) Solvent-accessible surface topology of a rhinovirus 14–drug complex (Kim *et al.*, 1993). The triangle shows one of the 60 symmetry-equivalent faces of an icosadeltahedron that constitute the entire virus surface. The surface is coloured and contoured according to distance from the centre of the virus, red being the most elevated. Residues are marked with dotted lines and labelled with residue type and number. A letter starting the residue label indicates a symmetry equivalent. The first numeral indicates the protein number (1 to 4), which is followed by the three-digit residue number. A depression, the 'canyon', is where the cellular receptor is bound (Olson *et al.*, 1993). The locations of the dominant neutralizing immunogenic (NIm) sites were determined through the sequencing of escape mutants (Sherry & Rueckert, 1985; Sherry *et al.*, 1986) and are labelled 'NIm'. (*b*) The same view is coloured according to sequence similarity (Palmenberg, 1989; Chapman, 1994), with blue being the most conserved rhinoviral amino acids and red being the most variable. Comparison of diagrams like these suggested the 'canyon hypothesis' (Rossmann, 1989). The prediction has proved largely true in that the sites of receptor attachment in several picornaviruses would be depressed areas whose sequences could be more highly conserved because they were partially inaccessible to antibodies and therefore not under the same selective pressure to mutate. In this and other applications, the schematic nature of these diagrams has helped in the collation of structure with data arising from the known phenotypes of site-directed or natural mutants. Part (*b*) is reproduced from Chapman (1993). Copyright (1993) The Protein Society. Reprinted with the permission of Cambridge University Press.

Are there similarities in the shapes of surfaces at the interfaces of macromolecular complexes? For example, are there similarities between the shapes of evolutionary-related antigens or the hypervariable regions of antibodies that bind to them? Quantitative comparison of surface topologies is far from trivial, with questions of three-dimensional (3D) alignment, the metrics to be used in quantifying topology *etc.* In addition to real differences between molecules, their surfaces may appear to differ due to the resolutions at which their structures were determined. Gerstein (1992) has proposed that comparisons be made in reciprocal space so that correlations can be judged as a function of resolution. Coordinates are aligned. Spherical Gaussian functions are placed at each atom, and an envelope is calculated at some threshold value and modified to remove cavities. Gerstein found that comparison of the envelope structure-factor vectors, obtained by Fourier transformation, led to a plausible classification of the hypervariable regions of known antibody structures.

22.2.4. Graphical representations of shape and properties

22.2.4.1. *Realistic*

22.2.4.1.1. *Shaded backbone*

With very large complexes, such as viruses, the surface features to be viewed are obvious at low resolution. In a very simple yet effective representation popularized by the laboratories of David Stuart and Jim Hogle, a Cα trace is 'depth cued' (shaded) according to the distance from the centre of mass (Acharya *et al.*, 1990; Fig. 1 for example). The impression of three dimensions probably results from the similarity of the shading to highlighting. The method is most effective for large complexes in which there are sufficient Cα atoms to give a dense impression of a surface.

22.2.4.1.2. *'Connolly' and solid polyhedral surfaces*

In one of the earliest surface graphical representations, dots were drawn for each Connolly surface dot, using vector-graphics terminals. With the improved graphics capability of modern computers, dot representations have been replaced by ones in which solid polyhedra are drawn with a large enough number of small triangular faces such that the surface appears smooth. These representations are clearer, because atoms in the foreground obscure those in the background.

22.2.4.1.3. *Photorealistic rendering*

Depth and three-dimensional relationships are most easily represented by stereovision or rotation of objects in real time on a computer screen. Graphics engines for interactive computers compromise quality for the speed necessary for interactive response, but simple depth cueing (combined with motion or stereo) is sufficient for good 3D representation. For still and/or non-stereo images more common in publications, more sophisticated rendering is helpful and possible now that speed is not a constraint. In *Raster3D* (Merritt & Bacon, 1997), multiple-light-source shading and highlighting is added, with individual calculations for each fine pixel. These are dependent on the directions

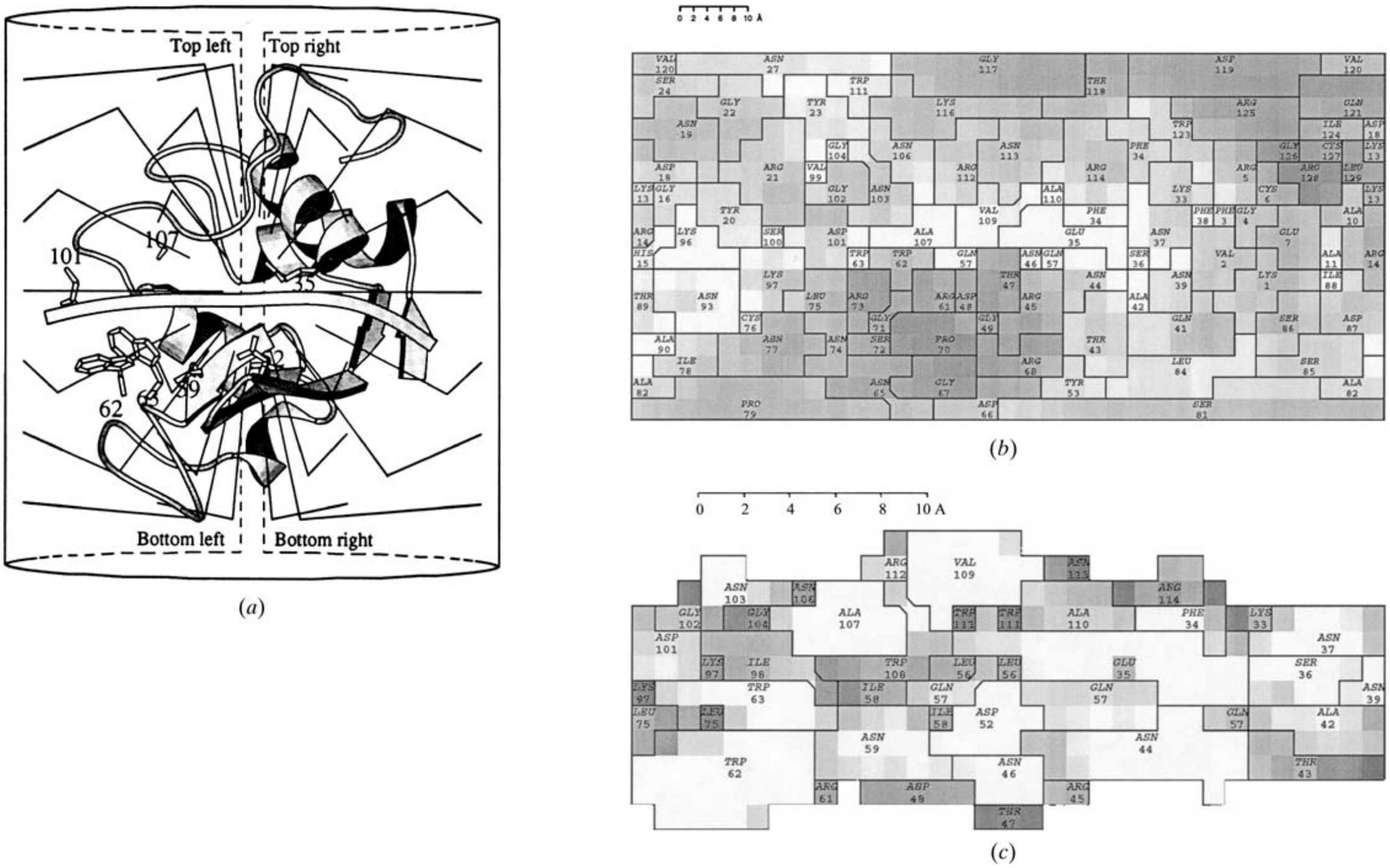


Figure 22.2.4.3

Different projections illustrated with lysozyme. (*a*) Lysozyme (Blake *et al.*, 1965; Diamond, 1974) is sketched with *MOLSCRIPT* (Kraulis, 1991) and shown with a ribbon indicting the active-site cleft (Kelly *et al.*, 1979). A cylindrical surface shown unrolled in (*b*) is shown in (*a*) wrapped around lysozyme. Vectors orthogonal to the now cylindrical illustrative surface are extended inwards until they intersect with the sphere. Vectors then run towards the centre of the molecule, and their intersections with the solvent-accessible surface are projected back upon the cylinder [unrolled in (*b*)]. (*b*) The surface is shaded according to distance from the centre, revealing the substrate-binding cleft as lighter shading. Details of active-site residues are revealed in (*c*) with a different type of projection. A segmented bent cylinder was traced along the substrate-binding cleft. The surface shows the projection outwards from points on the cylindrical axis. This reveals the amino acids likely to be in most intimate contact with the substrate. Similar plots, coloured according to charge, atomic solvation parameters, or hydrophobicity, can reveal the nature of predominant chemical interactions. This figure is reproduced from Chapman (1993). Copyright (1993) The Protein Society. Reprinted with the permission of Cambridge University Press.

of the normals to the surface, which are calculated analytically for spherical surfaces. More complicated surfaces, input as connected triangles, have surfaces rendered raster, pixel by pixel, by interpolating between the surface-normal vectors at the vertices of the surrounding triangle. Together, this leads to a high-quality smooth image that conveys much of the three-dimensionality of molecular surfaces.

22.2.4.1.4. GRASP surfaces

GRASP is currently perhaps the most popular program for the display of molecular surfaces. Readers are referred to the program documentation (Nicholls, 1992) or a paper that tangentially describes an early implementation (Nicholls *et al.*, 1991). The molecular or accessible surface is determined by the marching-cube algorithm. The surface is filled using methods that make modest compromises on photorealistic light reflection *etc.*, but take advantage of machine-dependent Silicon Graphics surface rendering to perform the display fast enough for interactive adjustment of the view.

The most powerful part of the program is the ability to colour according to properties mapped to the surface (see Fig. 22.2.4.1). These may be values of (say) electrostatic potential interpolated from a three-dimensional lattice. Much has been learned about many proteins from the potentials determined by solution of the Poisson–Boltzmann equation (Nicholls & Honig, 1991). The electrostatic complementarity of binding surfaces has often been readily apparent in ways that were not obvious from Coulombic calculations that ignore screening or from calculations and graphics representations that treat the charges of individual atoms as independent entities.

Many other properties can be mapped to the surface. These include properties of the atoms associated with that part of the surface (such as thermal factors), curvature of the surface calculated from adjacent atoms (Nicholls & Honig, 1991), or distance to the nearest part of the surface of an adjacent molecule. *GRASP* is now used to illustrate complicated molecular structures, in part because it also supports the superimposition of other objects over the molecular surface. These include the representation of molecules with Corey–Pauling–Koltun (CPK) spheres and/or bonds, and the representation of electrostatic potentials with field lines, dipole vectors *etc.*

22.2.4.1.5. Implementations in popular packages

Commercial packages use variants of the methods discussed above. For example, surfaces are drawn in the *Insight II* molecular modelling system using the Connolly dot algorithm (Molecular Structure Corporation, 1995).

22.2.4.2. *Schematic and two-dimensional representations such as 'roadmap'*

For their work on viruses, Rossmann & Palmenberg (1988) introduced a highly schematic representation in which individual amino acids were labelled. The methods were extended by Chapman (1993) to other proteins and to the automatic display of features such as topology, sequence similarity and hydrophobicity. Roadmaps sacrifice a realistic impression of shape for the ability to show the locations and properties of constituent surface atoms or residues. This has been important in combining the power of structure and molecular biology in understanding function. Potential sites of mutation are readily identified without substantial molecular-graphics resources, and phenotypes of mutants are readily mapped to the surface and compared with the physiochemical properties to reveal structure–function correlations.

For a set of projection vectors, the intersection points with the first van der Waals (or solvent-accessible) surface of an atom are calculated by basic vector algebra. The atom is identified so that when the projection is mapped to a plane for display, the boundaries of each atom or amino acid can be determined. The atoms or amino acids can then be coloured, shaded, outlined, contoured, or labelled according to parameters that are either calculated from the coordinates (such as distance from the centre of mass), read from a file (such as sequence similarity), or follow properties that are dependent on the residue type (*e.g.* hydrophobicity) or atom type [*e.g.* atomic solvation parameters (Eisenberg & McLachlan, 1986)].

Several types of projections can be used. The simplest is similar to that used by most other surface-imaging programs. A set of parallel projection vectors is mapped to a 2D grid. An example is shown in Fig. 22.2.4.2. This view avoids distortions, but only one side of the molecule is visualized. Roadmaps are flat, two-dimensional projections that cannot be rotated in real time to reveal other views. Three-dimensionality is limited to an extension by Jean-Yves Sgro that maps the parallel projection of one view to a three-dimensional surface shell that can be rotated with interactive graphics and/or viewed with stereo imaging (Harber *et al.*, 1995; Sgro, 1996). However, the schematic nature of roadmaps leads to the ability to view all parts of the molecule simultaneously.

To view all parts of the molecule, cylindrical projections are used that are similar to those used in atlases. This is possible because the representation is schematic (not realistic), and longitudinal distortion, similar to that near the poles in world maps, is acceptable. The surface is projected outwards radially onto a cylinder that wraps around the macromolecule (Fig. 22.2.4.3). Active-site clefts, drug or inhibitor binding sites and pores can be similarly illustrated by projecting their surfaces outward (from the axis) onto a cylinder that encloses the pore, pocket, or cleft. Such clefts are rarely straight, but with some distortion a satisfactory representation is possible by segmenting the cylinder, so that its axis follows the (curved) centre of the binding site or pore (Fig. 22.2.4.3).

22.2.5. Conclusion

Both quantitative and qualitative analyses of the surfaces of biomolecules are among the most powerful methods of elucidating functional mechanism from three-dimensional structures. A wide array of methods have been developed to help understand binding interactions and macromolecular assembly and to visualize the shape and physiochemical surface properties of macromolecules. Visualization methods range from those that depict a realistic impression of the topology to those that are more schematic and facilitate collation of structural and genetic information.

The authors thank Genfa Zhou for providing Fig. 22.2.4.1. MSC gratefully acknowledges the support of the National Science Foundation (BIR 94-18741 and DBI 98-08098), the National Institutes of Health (GM 55837) and the Markey Charitable Trust.

References

Acharya, R., Fry, E., Logan, D., Stuart, D., Brown, F., Fox, G. & Rowlands, D. (1990). *The three-dimensional structure of foot-and-mouth disease virus. New Aspects of Positive-Strand RNA Viruses*, edited by M. A. Brinton & S. X. Heinz, pp. 319–327. Washington DC: American Society for Microbiology.

Arnold, E. & Rossmann, M. G. (1990). *Analysis of the structure of a common cold virus, human rhinovirus 14, refined at a resolution of 3.0 Å. J. Mol. Biol.* **211**, 763–801.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme, a three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Chapman, M. S. (1993). *Mapping the surface properties of macromolecules. Protein Sci.* **2**, 459–469.

Chapman, M. S. (1994). *Sequence similarity scores and the inference of structure/function relationships. Comput. Appl. Biosci. (CABIOS)*, **10**, 111–119.

Chothia, C. (1976). *The nature of the accessible and buried surfaces in proteins. J. Mol. Biol.* **105**, 1–12.

Connolly, M. L. (1983). *Analytical molecular surface calculation. J. Appl. Cryst.* **16**, 548–558.

Diamond, R. (1974). *Real-space refinement of the structure of hen egg-white lysozyme. J. Mol. Biol.* **82**, 371–391.

Eisenberg, D. & McLachlan, A. D. (1986). *Solvation energy in protein folding and binding. Nature (London)*, **319**, 199–203.

Fauchere, J.-L. & Pliska, V. (1983). *Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. Eur. J. Med. Chem. Chim. Ther.* **18**, 369–375.

Fritz-Wolf, K., Schnyder, T., Wallimann, T. & Kabsch, W. (1996). *Structure of mitochondrial creatine kinase. Nature (London)*, **381**, 341–345.

Gerstein, M. (1992). *A resolution-sensitive procedure for comparing surfaces and its application to the comparison of antigen-combining sites. Acta Cryst.* A**48**, 271–276.

Gerstein, M. & Lynden-Bell, R. M. (1993). *What is the natural boundary for a protein in solution? J. Mol. Biol.* **230**, 641–650.

Grant, J. A. & Pickup, B. T. (1995). *A Gaussian description of molecular shape. J. Phys. Chem.* **99**, 3503–3510.

Greer, J. & Bush, B. L. (1978). *Macromolecular shape and surface maps by solvent exclusion. Proc. Natl Acad. Sci. USA*, **75**, 303–307.

Harber, J., Bernhardt, G., Lu, H.-H., Sgro, J.-Y. & Wimmer, E. (1995). *Canyon rim residues, including antigenic determinants, modulate serotype-specific binding of polioviruses to mutants of the poliovirus receptor. Virology*, **214**, 559–570.

Hermann, R. B. (1977). *Use of solvent cavity area and number of packed solvent molecules around a solute in regard to hydrocarbon solubilities and hydrophobic interactions. Proc. Natl Acad. Sci. USA*, **74**, 4144–4195.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–63.

Kelly, J. A., Sielecki, A. R., Sykes, B. D., James, M. N. & Phillips, D. C. (1979). *X-ray crystallography of the binding of the bacterial cell wall trisaccharaide NAM-NAG-NAM to lysozymes. Nature (London)*, **282**, 875–878.

Kim, K. H., Willingmann, P., Gong, Z. X., Kremer, M. J., Chapman, M. S., Minor, I., Oliveira, M. A., Rossmann, M. G., Andries, K., Diana, G. D., Dutko, F. J., McKinlay, M. A. & Pevear, D. C. (1993). *A comparison of the anti-rhinoviral drug binding pocket in HRV14 and HRV1A. J. Mol. Biol.* **230**, 206–227.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst.* **24**, 946–950.

Lee, B. & Richards, F. M. (1971). *The interpretation of protein structures: estimation of static accessibility*. *J. Mol. Biol.* **55**, 379–400.

Lewis, M. & Rees, D. C. (1985). *Fractal surfaces of proteins*. *Science*, **230**, 1163–1165.

Merritt, E. A. & Bacon, D. J. (1997). *Raster3D: photorealistic molecular graphics. Methods Enzymol.* **277**, 505–525.

Molecular Structure Corporation (1995). *Insight II User Guide*. Biosym/MSI, San Diego.

Nicholls, A. (1992). *GRASP: Graphical Representation and Analysis of Surface Properties.* New York: Columbia University.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Nicholls, A., Sharp, K. & Honig, B. (1991). *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins*, **11**, 281–296.

Olson, N., Kolatkar, P., Oliveira, M. A., Cheng, R. H., Greve, J. M., McClelland, A., Baker, T. S. & Rossmann, M. G. (1993). *Structure of a human rhinovirus complexed with its receptor molecule. Proc. Natl Acad. Sci. USA*, **90**, 507–511.

Palmenberg, A. C. (1989). *Sequence alignments of picornaviral capsid proteins*. In *Molecular Aspects of Picornavirus Infection and Detection*, edited by B. L. Semler & E. Ehrenfeld, pp. 211–241. Washington DC: American Society for Microbiology.

Pattabiraman, N., Ward, K. B. & Fleming, P. J. (1995). *Occluded molecular surface: analysis of protein packing*. *J. Mol. Recognit.* **8**, 334–344.

Reynolds, J. A., Gilbert, D. B. & Tanford, C. (1974). *Empirical correlation between hydrophobic free energy and aqueous cavity surface area*. *Proc. Natl Acad. Sci. USA*, **71**, 2925–2927.

Richards, F. M. (1977). *Areas, volumes, packing, and protein structure*. *Annu. Rev. Biophys. Bioeng.* **6**, 151–176.

Richards, F. M. (1985). *Calculation of molecular volumes and areas for structures of known geometry*. *Methods Enzymol.* **115**, 440–464.

Richmond, T. J. (1984). *Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. J. Mol. Biol.* **178**, 63–89.

Rossmann, M. G. (1989). *The canyon hypothesis. J. Biol. Chem.* **264**, 14587–14590.

Rossmann, M. G. & Palmenberg, A. C. (1988). *Conservation of the putative receptor attachment site in picornaviruses. Virology*, **164**, 373–382.

Sgro, J.-Y. (1996). *Virus visualization*. In *Encyclopedia of Virology Plus* (CD-ROM version), edited by R. G. Webster & A. Granoff. San Diego: Academic Press.

Sharp, K. A., Nicholls, A., Fine, R. F. & Honig, B. (1991). *Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science*, **252**, 107–109.

Sherry, B., Mosser, A. G., Colonno, R. J. & Rueckert, R. R. (1986). *Use of monoclonal antibodies to identify four neutralization immunogens on a common cold picornavirus, human rhinovirus 14. J. Virol.* **57**, 246–257.

Sherry, B. & Rueckert, R. (1985). *Evidence for at least two dominant neutralization antigens on human rhinovirus 14. J. Virol.* **53**, 137–143.

Tanford, C. (1997). *How protein chemists learned about the hydrophobicity factor. Protein Sci.* **6**, 1358–1366.

Tanford, C. H. (1979). *Interfacial free energy and the hydrophobic effect*. *Proc. Natl Acad. Sci. USA*, **76**, 4175–4176.

Ten Eyck, L. F. (1977). *Efficient structure-factor calculation for large molecules by the fast Fourier transform. Acta Cryst.* A**33**, 486–492.

Wodak, S. J. & Janin, J. (1980). *Analytical approximation to the accessible surface areas of proteins. Proc. Natl Acad. Sci. USA*, **77**, 1736–1740.

Xie, Q. & Chapman, M. S. (1996). *Canine parvovirus capsid structure, analyzed at 2.9 Å resolution. J. Mol. Biol.* **264**, 497–520.

Zhou, G., Somasundaram, T., Blanc, E., Parthasarathy, G., Ellington, W. R. & Chapman, M. S. (1998). *Transition state structure of arginine kinase: implications for catalysis of bimolecular reactions. Proc. Natl Acad. Sci. USA*, **95**, 8449–8454.

International Tables for Crystallography (2012). Vol. F, Chapter 22.3, pp. 721–729.

Chapter 22.3. Hydrogen bonding in biological macromolecules

E. N. Baker

22.3.1. Introduction

The hydrogen bond (Huggins, 1971) plays a critical role in the structure and function of biological macromolecules. This is because, uniquely among the non-covalent interactions that stabilize such structures, it combines a strong directional character with its energetic contributions. Thus, hydrogen-bonding patterns define the secondary structures that form the framework of proteins, are responsible for the specificity of base pairing in nucleic acids, shape the loops and irregular features that often determine molecular recognition, and provide for appropriately oriented functional groups in catalytic and/or binding sites.

Much of our present knowledge of hydrogen bonding in biological structures is foreshadowed in Linus Pauling's influential book (Pauling, 1960), and Jeffrey & Saenger (1991) provide a very comprehensive review. Other important reviews have covered hydrogen-bonding patterns in globular proteins (Baker & Hubbard, 1984; Stickle *et al.*, 1992), and more focused reviews address particular aspects: the satisfaction of hydrogen-bonding potential in proteins (McDonald & Thornton, 1994*a*), hydrogen-bonding patterns for side chains (Ippolito *et al.*, 1990), side-chain hydrogen bonding in relation to secondary structures (Bordo & Argos, 1994), distributions of side-chain–main-chain hydrogen bonds (Eswar & Ramakrishnan, 2000) and the hydrogen bonds associated with protein–ligand interactions (Torrance *et al.*, 2007; Liu *et al.*, 2008). The prevalence and importance of weak and non-conventional hydrogen bonds has been comprehensively reviewed by Desiraju & Steiner (1999).

22.3.2. Nature of the hydrogen bond

Hydrogen bonds have traditionally been defined as attractive electrostatic interactions of the type $D-\mathrm{H}\cdots A$, where the H atom is formally attached to a donor atom, D (assumed to be more negative than H), and is directed towards an electron-rich acceptor, A. The acceptor A is normally an electronegative atom, usually O or N, but occasionally S or Cl, with a full or partial negative charge and a lone pair of electrons directed towards the H atom. Although most of the hydrogen bonds in proteins and nucleic acids are $\mathrm{N-H\cdots O}$ or $\mathrm{O-H\cdots O}$ (less often, $\mathrm{N-H\cdots N}$), it is important to be aware that other possibilities exist, including $\mathrm{N-H\cdots S}$, $\mathrm{O-H\cdots S}$ and $\mathrm{C-H\cdots O}$, and that these can be very important in specific cases (Adman *et al.*, 1975; Derewenda *et al.*, 1995). Likewise, the π-electron clouds of aromatic rings can also act as acceptors for appropriately oriented $D-\mathrm{H}$ groups (Legon & Millen, 1987; Mitchell *et al.*, 1994). To cover this full spectrum of interactions, the formal definition of the hydrogen bond has recently been revised; see the definition in Desiraju (2011).

In an ideal hydrogen bond, the donor heavy atom, the H atom, the acceptor lone pair and the acceptor heavy atom should all lie in a straight line (Legon & Millen, 1987), as illustrated in Fig. 22.3.2.1(*a*). The strength of the interaction is also expected to depend on the electronegativities of the atoms involved. Hydrogen bonds are said to be bifurcated when a single $D-\mathrm{H}$ group interacts with two acceptors in a three-centred hydrogen bond (Fig. 22.3.2.1*b*); these hydrogen bonds are necessarily nonlinear and weaker. However, the term bifurcated is also sometimes applied to the quite different situation where a donor atom with two H atoms or an acceptor atom with two lone pairs makes two hydrogen bonds, as in Figs. 22.3.2.1(*c*) and (*d*). These interactions can be strong and linear. Some hydrogen-bonding arrangements are said to be cooperative; for example, hydrogen bonding by a peptide C═O group should enhance the polarity of the whole peptide unit and hence the acidity of the amide proton and the strength of its hydrogen bonding (Jeffrey & Saenger, 1991).

22.3.3. Hydrogen-bonding groups

22.3.3.1. *Proteins*

The hydrogen-bonding capacities of the various hydrogen-bonding groups in proteins are shown in Fig. 22.3.3.1. All, with the exception of the peptide NH and Trp side-chain NH groups, can participate in more than one hydrogen-bond interaction. Peptide and side-chain C═O groups, for example, can act as acceptors for two hydrogen bonds by using both lone pairs of electrons on the sp^2-hybridized oxygen. Likewise, the —OH groups of Ser or Thr can act as donors through their single H atom, and acceptors through their two lone pairs. In Tyr side chains, the C—O bond has some double-bond character, and the phenolic —OH is thus likely to prefer only two hydrogen bonds, both in the ring plane. The carboxylate groups of Asp and Glu are normally ionized above pH 4 and their C—O bonds also have partial double-bond character; each carboxylate oxygen should then be able to accept two hydrogen bonds, although the restriction to two may be less severe than for C═O.

Several uncertainties exist. Crystallographically, it is not usually possible to distinguish the amide oxygen and nitrogen atoms of Asn and Gln, and the decision as to which is which has to be made on environmental grounds by considering what

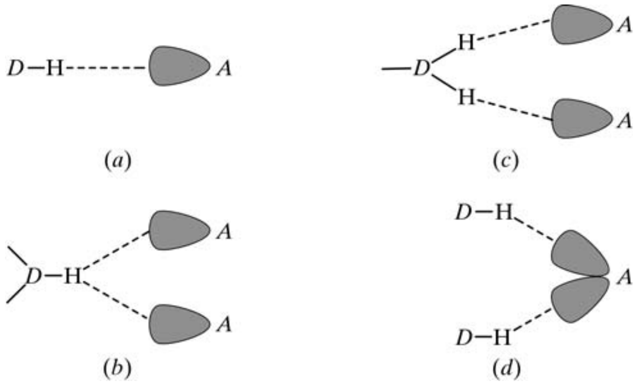


Figure 22.3.2.1
Hydrogen-bonding configurations. (*a*) The standard two-centre hydrogen bond in which an H atom attached to a donor atom, *D*, is directed towards a lone pair of an acceptor, *A*. (*b*) A classic three-centre, or bifurcated, hydrogen bond, with a single H atom shared between the lone pairs of two acceptors. The situations shown in (*c*) and (*d*) are not true three-centre hydrogen bonds since they are essentially equivalent to that in (*a*).

Copyright © 2012 International Union of Crystallography

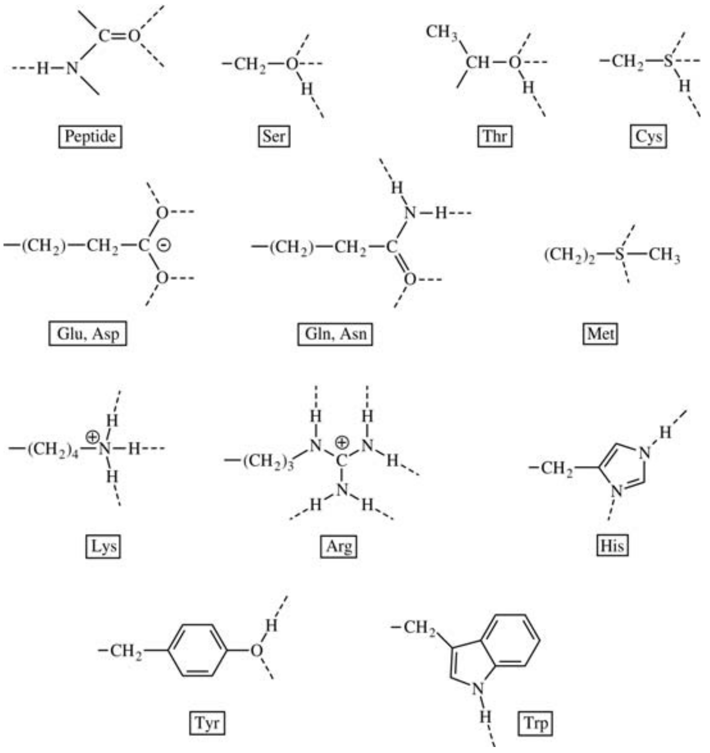


Figure 22.3.3.1
Hydrogen-bonding potential of protein functional groups. Potential hydrogen bonds are shown with broken lines. Arg, Lys, Asp and Glu side chains are shown in their ionized forms.

hydrogen bonds would be made in each of the two possible arrangements. Likewise, two possibilities exist for His side chains by rotating 180° about $C^{\beta}-C^{\gamma}$. This problem is discussed by McDonald & Thornton (1994*b*), and is addressed by their program *HBPLUS* and by validation programs such as *MolProbity* (Chapter 21.6; Richardson *et al.*, 2003).

For some side chains, the ionization state is uncertain. Arg and Lys are assumed to be fully protonated, as in Fig. 22.3.3.1, and Asp and Glu are assumed to be fully ionized. Nevertheless, several surveys of interactions in protein structures (Flocco & Mowbray, 1995; Langkilde *et al.*, 2008) show that a small but significant number of short O$\cdots$O distances between Asp and Glu side chains must represent O$-$H$\cdots$O hydrogen bonds, with one carboxyl group protonated. His side chains, in addition to the orientational uncertainty, have a pK_a (~6.5) that implies that they may be in either their neutral or their protonated form, depending on pH and environment. In the neutral form, only one N atom is protonated (more often $N^{\varepsilon 2}$, but sometimes $N^{\delta 1}$), but in the protonated form both N atoms carry protons; again, the actual state has to be deduced from their environment.

22.3.3.2. *Nucleic acids*

The three components of nucleic acids, *i.e.* phosphate groups, sugars and bases, all participate in hydrogen bonding to greater or lesser extent. The phosphate oxygen atoms can potentially act as acceptors of two or more hydrogen bonds and are frequently the recipients of hydrogen bonds from protein side chains in protein–DNA complexes. The sugar residues of RNA have a 2′-OH which can act as both hydrogen-bond donor and acceptor, and the 4′-O of both ribose and deoxyribose can potentially accept two hydrogen bonds.

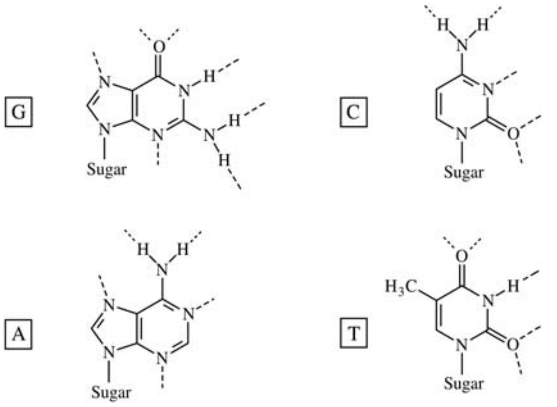


Figure 22.3.3.2
Hydrogen-bonding potential of nucleic acid bases guanine (G), adenine (A), cytosine (C) and thymine (T) in their normal canonical forms.

It is the bases of DNA and RNA that have the greatest hydrogen-bonding potential, however, with a variety of hydrogen-bond donor or acceptor sites. Although each of the bases could theoretically occur in several tautomeric forms, only the canonical forms shown in Fig. 22.3.3.2 are actually observed in nucleic acids. This leads to clearly defined hydrogen-bonding patterns which are critical to both base pairing and protein–nucleic acid recognition. The $-NH_2$ and >NH groups act only as hydrogen-bond donors, and C═O only as acceptors, whereas the >N$-$ centres are normally acceptors but at low pH can be protonated and act as hydrogen-bond donors.

22.3.4. Identification of hydrogen bonds: geometrical considerations

Because hydrogen bonds are electrostatic interactions for which the attractive energy falls off rather slowly (Hagler *et al.*, 1974), it is not possible to choose an exact cutoff for hydrogen-bonding distances. Rather, both distances and angles must be considered together; the latter are particularly important because of the directionality of hydrogen bonding. Inferences drawn from distances alone can be highly misleading. An approach with an N$-$H$\cdots$O angle of 90° and an H$\cdots$O distance of 2.5 Å would be very unfavourable for hydrogen bonding, yet it translates to a N$\cdots$O distance of 2.7 Å. This could (wrongly) be taken as evidence of a strong hydrogen bond.

For macromolecular structures determined by X-ray crystallography, problems also arise from the imprecision of atomic positions and the fact that H atoms cannot usually be seen. Thus, the geometric criteria must be relatively liberal. H atoms should also be added in calculated positions where this is possible; this can be done reliably for most NH groups (peptide NH, side chains of Trp, Asn, Gln, Arg, His, and all >NH and NH_2 groups in nucleic acid bases).

The hydrogen-bond criteria used by Baker & Hubbard (1984) are shown in Fig. 22.3.4.1. Very similar criteria are used in the program *HBPLUS* (McDonald & Thornton, 1994*a*), which also adds H atoms in their calculated positions if they are not already present in the coordinate file. In general, hydrogen bonds may be inferred if an interatomic contact obeys *all* of the following criteria:

(1) The distance H$\cdots$$A$ is less than 2.5 Å (or $D\cdots A$ less than 3.5 Å if the donor is an $-$OH or $-NH_3^+$ group or a water molecule).
(2) The angle at the H atom, $D-\mathrm{H}\cdots A$, is greater than 90°.
(3) The angle at the acceptor, $AA-A\cdots\mathrm{H}$ (or $AA-A\cdots D$ if the H-atom position is unreliable), is greater than 90°.

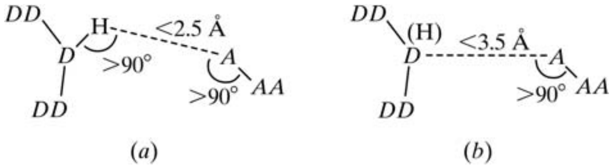


Figure 22.3.4.1
Suggested criteria for identifying likely hydrogen bonds. *DD* and *AA* represent atoms covalently bonded to the donor atom, *D*, and acceptor atom, *A*, respectively. Here, (*a*) represents the criteria when the donor H atom can be placed, and (*b*) when it cannot be placed. Additional criteria based on the angle $DD-D\cdots A$ could be incorporated with (*b*). Adapted from Baker & Hubbard (1984) and McDonald & Thornton (1994*a*).

Other criteria can be applied, for example taking into account the hybridization state of the atoms involved and the degree to which any approach lies in the plane of the lone pair(s). In all analyses of hydrogen bonding, however, it is clear that a combination of distance and angle criteria is effective in excluding unlikely hydrogen bonds.

22.3.5. Hydrogen bonding in proteins

22.3.5.1. Contribution to protein folding, stability and function

The net contribution of hydrogen bonding to protein folding and stability has been the subject of much debate over the years (Pace, 2009). The current view is that although the hydrophobic effect provides the driving force for protein folding (Kauzmann, 1959), many polar groups, notably peptide NH and C═O groups, inevitably become buried during this process, and failure of these groups to find hydrogen-bonding partners in the folded protein would be strongly destabilizing. This, therefore, favours the formation of secondary structures and other structures that permit effective hydrogen bonding in the folded molecule. Not surprisingly, the contribution of specific hydrogen bonds to stability depends on their location in the structure (Fersht & Serrano, 1993). Mutagenesis studies have shown that even the loss of a single hydrogen bond can be significantly destabilizing (Alber *et al.*, 1987) and that the energetic contribution can vary depending on whether or not the groups involved are charged (Fersht *et al.*, 1985). A recent experimental study shows that the polarity of the microenvironment of a hydrogen bond can increase its strength by up to 1.2 kcal mol^{-1} (1 kcal mol^{-1} = 4.184 kJ mol^{-1}) (Gao *et al.*, 2009).

Hydrogen bonds are fundamental to protein function, notably in the catalytic activity and specificity of enzymes (Gutteridge & Thornton, 2005). The concept of very short, low-barrier hydrogen bonds (LBHBs) that play a special role in enzyme catalysis has been proposed (Cleland *et al.*, 1998) but is controversial (Schutz & Warshel, 2004). Short, strong hydrogen bonds can be detected by NMR (see the discussion in Langkilde *et al.*, 2008) but do not stand out in crystal structures, and their functional role remains unclear.

22.3.5.2. Saturation of hydrogen-bond potential

A consistent conclusion from analyses of protein structures is that virtually all polar atoms either form explicit hydrogen bonds or are at least in contact with external water. The extent to which their *full* hydrogen-bond potential is fulfilled in a folded protein (for example, the potential of an Arg side chain to make five hydrogen bonds) has been examined in several studies. Baker & Hubbard (1984) considered the explicit hydrogen bonds made by main-chain and side-chain atoms in a number of refined protein structures and established general patterns for both, but did not differentiate buried and solvent-exposed atoms or allow for unmodelled solvent. Savage *et al.* (1993) used the solvent accessibilities of polar groups to estimate their assumed numbers of hydrogen bonds to external water. This supplemented the explicit hydrogen bonds that could be derived from the atomic coordinates and allowed an estimate of the extent to which potential hydrogen bonds are lost during protein folding. McDonald & Thornton (1994*a*) focused specifically on buried hydrogen-bond donors and acceptors in order to determine the extent to which the hydrogen-bond potential of these is utilized.

The results of these analyses can be summarized as follows. Almost all polar groups do in fact make at least one hydrogen bond. Hydrogen-bond donors are almost always hydrogen bonded; only 4% of NH groups 'lose' hydrogen bonds as a result of protein folding (Savage *et al.*, 1993). On the other hand, hydrogen-bond acceptors often do not exert their full hydrogen-bonding potential. For example, for main-chain C═O groups, which are expected to accept two hydrogen bonds, 24% of possible hydrogen bonds are estimated to be lost during folding (Savage *et al.*, 1993). Among buried C═O groups, although very few make no hydrogen bonds (as little as 2% if hydrogen-bonding criteria are relaxed), the majority fail to form a second hydrogen bond (McDonald & Thornton, 1994*a*). Steric factors, particularly in β-sheets or where Pro residues are adjacent, restrict hydrogen-bonding possibilities, although some of the 'lost' interactions may be recovered through $\mathrm{C-H\cdots O}$ interactions (see Section 22.3.7.1). McDonald & Thornton also point out that failure to form a second hydrogen bond is less energetically expensive than failure to form the first. Among polar side chains, the ionizable side chains (Asp, Glu, Arg, Lys, His) show a very strong tendency to be fully hydrogen bonded or solvent exposed. Buried Arg side chains, for example, frequently form all five possible hydrogen bonds. The side chains that most often fail to fulfil their full hydrogen-bond potential are Ser, Thr and Tyr; these almost always donate one hydrogen bond but frequently fail to accept one.

22.3.5.3. Secondary structures

Secondary structures provide the means whereby the polar C═O and NH groups of the polypeptide chain can remain effectively hydrogen bonded when they are buried within a folded globular protein. In doing so, they provide the framework of folding patterns and account for the majority of hydrogen bonds within protein structures. The three secondary-structure classes (helices, β-sheets and turns) are each characterized by specific hydrogen-bonding patterns, which can be used for objective identification of these structures (Stickle *et al.*, 1992).

22.3.5.3.1. Helices

Helices have traditionally been defined in terms of their $\mathrm{N-H\cdots O{=}C}$ hydrogen-bonding patterns as α-helices ($i \to i-4$), 3_{10}-helices ($i \to i-3$) or π-helices ($i \to i-5$); in an α-helix, for example, the peptide NH of residue 5 hydrogen bonds to the C═O of residue 1. In fact, the vast majority of helices in proteins are α-helices; 3_{10}-helices are rarely more than two turns (six residues) in length, and only a very small number of discrete, but short, π-helices have been seen so far (Weaver, 2000).

The residues within helices have characteristic main-chain torsion angles, (φ, ψ), of around (−63°, −40°) that cause the C═O groups to tilt outwards by about 14° from the helix axis (Baker & Hubbard, 1984). This results in somewhat less linear hydrogen bonding than in the original Pauling model (Pauling *et al.*, 1951), with a degree of distortion towards 3_{10}-helix geometry. Thus, weak $i \rightarrow i-3$ interactions are often made in addition to the more favourable $i \rightarrow i-4$ hydrogen bonds, giving hydrogen-bond networks that may enhance helix elasticity (Stickle *et al.*, 1992). Tilting outwards also makes the C=O groups more accessible for additional hydrogen bonds from side chains or water molecules. For the α-type, $i \rightarrow i-4$ interactions, the hydrogen-bond angles at both donor and acceptor atoms are quite tightly clustered (N—H···O ~157° and C=O···H ~147°). The hydrogen-bond lengths in helices average 2.06 (16) Å (O···H) or 2.99 (14) Å (O···N) (Baker & Hubbard, 1984).

Few helices are regular throughout their length. Many are curved or kinked such that one side (often the outer, solvent-exposed side) of the helix is opened up a bit and has longer hydrogen bonds (Blundell *et al.*, 1983; Baker & Hubbard, 1984). The bends are often associated with additional hydrogen bonds from water molecules or side chains to C=O groups that are tilted out more than usual. Curved helices are normal in coiled-coil structures and can enable long helices to pack more effectively in globular structures. Sometimes a kink can be functionally important, as in manganese superoxide dismutase, where a kink in a long helix, incorporating a π-type ($i \rightarrow i-5$) hydrogen bond, enables the optimal positioning of active-site residues (Edwards *et al.*, 1998).

The beginnings and ends of helices are sites of hydrogen-bonding variations which can be seen as characteristic 'termination motifs'. At helix N-termini, 3_{10}-type $i \rightarrow i-3$ (or bifurcated $i \rightarrow i-3$ and $i \rightarrow i-4$) hydrogen bonds are often found. At C-termini, two common patterns occur. In one, labelled α_{C1} by Baker & Hubbard (1984), there is a transition from α-type, $i \rightarrow i-4$ to 3_{10}-type, $i \rightarrow i-3$ hydrogen bonding, often with genuine bifurcated hydrogen bonds, as in Fig. 22.3.2.1(*b*), at the transition point. The other, labelled α_{C2} (Baker & Hubbard, 1984) or referred to as the 'Schellman motif' (Schellman, 1980), has a π-type, $i \rightarrow i-5$ hydrogen bond coupled with a 3_{10}-type, $i-1 \rightarrow i-4$ hydrogen bond; residue $i-1$ has a left-handed α configuration and is often Gly. The beginnings and ends of helices are also the sites of specific side-chain hydrogen-bonding patterns, referred to as N-caps and C-caps (Presta & Rose, 1988; Richardson & Richardson, 1988); these are described below.

22.3.5.3.2. *β-sheets*

β-sheets consist of short strands of polypeptide (typically 5–7 residues) running parallel or antiparallel and cross-linked by N—H···O=C hydrogen bonds. Although the (φ, ψ) angles of residues within β-sheets can be quite variable, the hydrogen-bonding patterns within these segments tend to be quite regular, as in the original Pauling models (Pauling & Corey, 1951). Occasional β-bulges in the middle of β-strands can interrupt the hydrogen-bonding pattern (Richardson *et al.*, 1978), but otherwise disruptions occur only at the ends of strands. The hydrogen bonds in β-sheets appear to be slightly shorter than those in helices, by ~0.1 Å, and also more linear (N—H···O ~ 160°, compared with ~157° in helices) (Baker & Hubbard, 1984). There also appears to be no difference between parallel and antiparallel β-sheets in the hydrogen-bond lengths and angles.

22.3.5.3.3. *Turns*

By far the most common type of turn is the β-turn, a sequence of four residues that brings about a reversal in the polypeptide chain direction. Hydrogen bonding does not seem to be essential for turn formation, but a common feature is a hydrogen bond between the C=O group of residue 1 and the NH group of residue 4, a 3_{10}-type, $i \rightarrow i-3$ interaction. Turns are also often associated with characteristic side-chain–main-chain hydrogen-bond configurations (see below). The hydrogen bonds in turns tend to be longer and less linear than those in helices and β-sheets; in particular, the angle at the acceptor oxygen atom C—O···H is around 120° (Baker & Hubbard, 1984).

In addition to β-turns, a small but significant number of γ-turns are found. In these three-residue turns, a hydrogen bond is formed between the C=O of residue 1 and the NH of residue 3, an $i \rightarrow i-2$ interaction. Although the approach to the acceptor oxygen atom is highly nonlinear (C—O···H ~ 100°), the nonlinearity at the H atom is less pronounced (N—H···O ~ 130–150°) (Baker & Hubbard, 1984). γ-turns are again of several types, depending on the configuration of the central residue. The classic γ-turn, first recognized by Matthews (1972) and Nemethy & Printz (1972), has a central residue with (φ, ψ) angles around (70°, −60°), which puts it in the normally disallowed region of the Ramachandran plot. More common, however, are structures in which an $i \rightarrow i-2$ hydrogen bond is associated with a central residue with a configuration around (90°, −70°) (Baker & Hubbard, 1984); these structures are not necessarily true turns in the sense of bringing about a sharp chain reversal, however.

22.3.5.3.4. *Aspects of in-plane geometry*

For hydrogen bonds involving sp^2 donors and/or acceptors, optimal interaction is expected to occur when the donor D—H group and the acceptor lone-pair orbital are coplanar (Taylor *et al.*, 1983). Analysis of 'in-plane' and 'out-of-plane' components of N—H···O hydrogen bonds in proteins shows that these have characteristic values for different secondary structures (Artymiuk & Blake, 1981; Baker & Hubbard, 1984). The out-of-plane component is tightly clustered at ~25° for helices and ~60° for the most common β-turns (type I and type III), but is widely scattered around a mean of 0° for β-sheets. The latter reflects different twists or curvature of β-sheets. The large out-of-plane component for turns is consistent with a relatively weak interaction.

22.3.5.4. *Side-chain hydrogen bonding*

An important concept in understanding the patterns of side-chain hydrogen bonding in proteins is that of local *versus* non-local interactions; local means that a side chain hydrogen bonds to another residue that is relatively close to it in the linear amino-acid sequence. Baker & Hubbard (1984) were first to introduce this distinction, with local defined as ±4 residues. Eswar & Ramakrishnan (2000) define local as ±5 residues, Bordo & Argos (1994) as ±6 residues and Stickle *et al.* (1992) as ±10 residues. The distinction is not important, but the distributions in all three analyses show that ±5 would encompass all the significant populations of local hydrogen bonds. Local hydrogen bonds, in which side chains interact with nearby main-chain atoms or other side chains, are evidently critical for protein folding. Non-local hydrogen bonds, although fewer in number (see below), in turn can be very important for stabilization of the folded protein.

If hydrogen bonds with water are excluded, a rule of thirds applies. Approximately one-third of the hydrogen bonds made by side chains (sch's) are with main-chain (mch) C=O groups, one-third are with main-chain NH groups, and one-third with other side chains. Within these populations, however, there are significant differences. For sch–mch(C=O) hydrogen bonds,

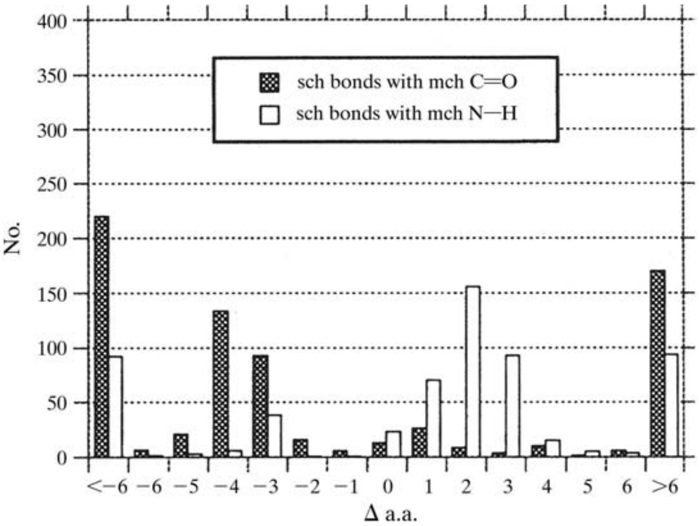


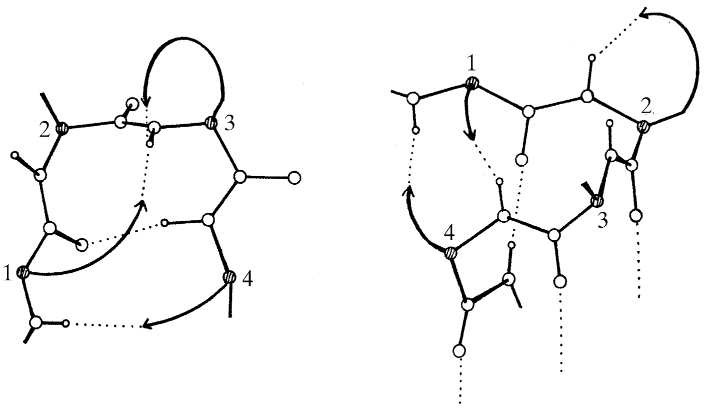
Figure 22.3.5.1
Distribution of side-chain–main-chain hydrogen bonds as a function of the separation (Δ a.a.) along the polypeptide between the side-chain (sch) and main-chain (mch) groups involved, *i.e.* Δ a.a. = −*n* means that a side chain interacts with a main-chain group *n* residues earlier in the polypeptide (towards the N-terminus). Reproduced with permission from Bordo & Argos (1994). Copyright (1994) Academic Press.

approximately 45% are local; for sch–mch(NH) hydrogen bonds, a much higher proportion is local (69%), and for sch–sch hydrogen bonds, the proportion is much less (35%) (Bordo & Argos, 1994).

The distribution of local sch–mch(NH) hydrogen bonds shows a marked positional preference (Fig. 22.3.5.1) that highlights consistent hydrogen-bonding motifs found in all proteins (Fig. 22.3.5.2). The major peak involves side chains that interact with an NH group two residues further on in the polypeptide, an *n*–NH(*n* + 2) hydrogen bond. This motif primarily involves Asp, Asn, Ser and Thr side chains and is most often found (i) in turns, where a side chain from position 1 hydrogen bonds to the NH of residue 3, (ii) in loop regions where it stabilizes the local structure but is not necessarily associated with chain reversal, and (iii) at helix N-termini.

Helix N-termini are also the site of other characteristic local side-chain–NH hydrogen-bonding motifs (Baker & Hubbard, 1984; Presta & Rose, 1988; Richardson & Richardson, 1988; Harper & Rose, 1993; Bordo & Argos, 1994). Prominent among these are sch–NH(*n* + 3) hydrogen bonds involving Ser, Thr, Asp and Asn side chains, but sch–NH(*n* − 3) interactions, in which Glu or Gln side chains hydrogen bond back to a main-chain NH, form an important lesser category. Other motifs, such as that in which a Glu or Gln side chain bends round to hydrogen bond to its own NH group, are also found. Collectively, these contribute to helix capping motifs (Fig. 22.3.5.2*b*) that help satisfy the hydrogen bonding of the 'free' NH groups of the helix N-terminus and in effect extend the helix; the sch–mch(NH) hydrogen bond mimics the mch–mch C═O···HN hydrogen bonds of the helix. Helix N-capping by side chains is probably a very important influence in protein folding, acting as a stereochemical code for helix initiation (Presta & Rose, 1988; Harper & Rose, 1993).

The distribution of sch–mch(CO) hydrogen bonds also shows a striking preference, this time for positions −3 and −4. These sch–CO(*n* − 3) or sch–CO(*n* − 4) hydrogen bonds account for the vast majority of local hydrogen bonds between side chains and main-chain C═O groups. Almost all (∼85%) are in helices, with most of the remainder in turns. They involve predominantly (∼80%) Ser and Thr side chains but other side chains (Asn, His, Arg) can also participate. These local hydrogen bonds can occur at any point along a helix, where they are often associated with helix bending or kinking (Baker & Hubbard, 1984). However, they are most frequently found at helix C-termini (Bordo & Argos, 1994) and may constitute a termination motif.



(*a*) (*b*)

Figure 22.3.5.2
Schematic representations of common classes of side-chain–main-chain hydrogen bonds (*a*) in turns and (*b*) at helix N-termini. Arrows represent side chains that hydrogen bond to main-chain CO or NH groups (NH identified by the small circle for H).

Local side-chain–side-chain hydrogen bonds, although common, do not seem to fit into any obvious patterns; the only recurring interaction identified so far is between side chains on succeeding turns of helices, *i.e.* separated by approximately four residues. These frequently involve charged side chains, which can form hydrogen-bonded ion pairs. In sections of extended chain, side chains that are two residues apart may similarly interact.

Non-local hydrogen bonding by side chains is less easy to categorize but is no less significant; more than 50% of side-chain–main-chain(C═O) hydrogen bonds are non-local, as are ∼65% of side-chain–side-chain hydrogen bonds. In most proteins, a small number of polar side chains with multiple hydrogen-bonding capability act as the centre for networks of hydrogen bonds; these appear to be particularly important for stabilizing non-repetitive polypeptide chain structures (coil, loops). Examples are given in Baker & Hubbard (1984). Most often these involve larger side chains with more than one hydrogen-bonding centre (Asn, Asp, Gln, Glu, Arg, His) which cross-link different sections of the polypeptide. Arg side chains interacting with main-chain C═O groups seem to be particularly effective; Ser and Thr, on the other hand, are seldom used, even though both have the potential to form three hydrogen bonds.

The geometry of side-chain hydrogen bonding has been analysed by Baker & Hubbard (1984) and, more extensively, by Ippolito *et al.* (1990). The former concentrate on hydrogen-bond lengths and angles and show that the preferred angles fit well with stereochemical expectations. Ippolito *et al.* examine the preferences for the various hydrogen-bonding sites around each side-chain type by means of scatter plots (Fig. 22.3.5.3) from which probability densities are computed. These show that well defined preferences exist, determined by both steric and electronic effects.

22.3.5.5. *Hydrogen bonds with water molecules*

Water molecules, with their small size and double-donor, double-acceptor hydrogen-bonding capability, are ideal for completing intramolecular hydrogen-bonding networks, *e.g.* by linking two proton acceptor atoms, or two protein donor atoms, that cannot otherwise interact. Thus, buried water molecules, making multiple hydrogen bonds, help satisfy the hydrogen-bond potential of internal polar atoms and contribute to protein stability; internal waters average about three hydrogen bonds each (Baker & Hubbard, 1984; Williams *et al.*, 1994). From the survey of Williams *et al.* (1994), most (58%) occupy discrete cavities, while 22% are in clusters housing two waters and 20% are in larger clusters; some examples of larger clusters are given in Baker & Hubbard (1984). Buried waters are often conserved between homologous proteins (Baker, 1995), and each buried water–protein hydrogen bond is estimated to stabilize a folded protein by, on average, 0.6 kcal mol^{-1} (Williams *et al.*, 1994). More loosely bound external waters exchange much more rapidly and presumably contribute less energetically.

Several patterns of hydrogen bonding are consistently observed. Water molecules are most often seen interacting with oxygen atoms rather than nitrogen atoms and acting as hydrogen-bond donors rather than acceptors. Possible reasons include the greater number of acceptor sites in proteins and the fewer geometrical restrictions imposed by acceptors (Baker & Hubbard, 1984; Baker, 1995). There is also a predominance of interactions with main-chain atoms rather than side-chain atoms: on average ~40% with main-chain C=O groups, 15% with main-chain NH and 45% with side-chain groups (Baker & Hubbard, 1984; Thanki *et al.*, 1988). Favoured main-chain binding sites include the N- and C-termini of helices, C=O groups on the solvent-exposed sides of helices, the edge strands of β-sheets, and the ends of strands where they add extra inter-strand hydrogen bonds at the position where the strands diverge (Thanki *et al.*, 1991). Among side chains, the most highly hydrated appear to be Asp and Glu, whose COO$^-$ groups bind, on average, two water molecules each (Baker & Hubbard, 1984; Thanki *et al.*, 1988). On the other hand, the best-ordered water sites are created by residues whose side chains simultaneously make hydrogen bonds to other protein atoms (His, Asp, Asn, Arg) or may be sterically restricted (Tyr, Trp).

The distributions of water molecules around protein groups follow the geometrical patterns expected from simple bonding ideas (Baker & Hubbard, 1984; Thanki *et al.*, 1988). Interactions with NH groups are linear, and those with C=O groups show a preferred angle of ~130° at the oxygen-atom acceptor, consistent with interaction with an oxygen-atom lone pair; restriction to the peptide plane is not very strong, however. Although the distributions around polar side chains generally follow the expected patterns (Thanki *et al.*, 1988), there is little evidence of ordered water clusters around non-polar groups. This may be because water clusters need to be 'anchored' by hydrogen bonding to polar groups to be seen crystallographically.

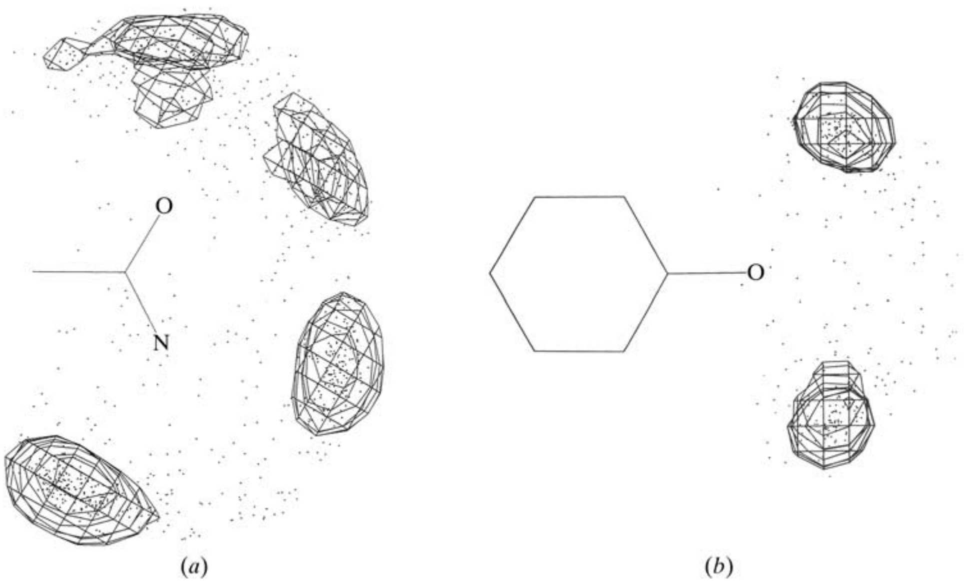


Figure 22.3.5.3
Typical scatter plots showing the distribution of hydrogen-bonding partners around protein side chains, shown for (*a*) Asn or Gln and (*b*) Tyr. Reproduced with permission from Ippolito *et al.* (1990). Copyright (1990) Academic Press.

22.3.6. Hydrogen bonding in nucleic acids

Hydrogen bonding by purine and pyrimidine bases is, together with base stacking, a major determinant of nucleic acid structure. With so many hydrogen-bonding groups, there are many potential modes of interaction between bases (Jeffrey & Saenger, 1991). Those that are actually found in DNA and RNA structures are, however, much more restricted in number, at least based on presently available experimental data.

22.3.6.1. DNA

DNA structure is dominated by the prevalence of duplex structures and hence by the classic Watson–Crick hydrogen-bonding pattern of A–T and G–C base pairs. This hydrogen-bonding pattern is not affected by whether the double helix has A-form, B-form or Z-form geometry. Other hydrogen-bonding modes in DNA are probably very rare, arising only as a result of mutations (which produce mismatches), chemical modifications, such as methylation, or other disturbances, such as the binding of drugs or proteins so as to alter DNA conformation. Mismatches can give stable hydrogen bonding but at the expense of local perturbations of the DNA structure.

22.3.6.2. RNA

In contrast to DNA, RNA molecules generally form single-stranded structures, which are correspondingly much more complex and less regular. This means that catalytic and other activities can be generated in addition to their information-carrying roles. Current knowledge of detailed RNA three-dimensional structure is limited to transfer RNAs and several ribozymes, including a large ribosomal RNA domain (Cate *et al.*, 1996). Even from this small sample, however, it is clear that a great diversity of hydrogen-bonding interactions exists; RNA molecules contain regions of double-helical structure, often with classical Watson–Crick A–U and G–C base pairing, but these regions are interspersed with loops and bulges and tertiary interactions between the various secondary-structural (double-helical) elements. These interactions include many unconventional base pairings (*e.g.* see Fig. 22.3.6.1).

Some RNA structural motifs may prove to be of widespread general importance in RNA molecules. One example is a sharp turn with sequence CUGA in the hammerhead ribozyme that exactly matches turns in tRNAs (Pley *et al.*, 1994). Another is the

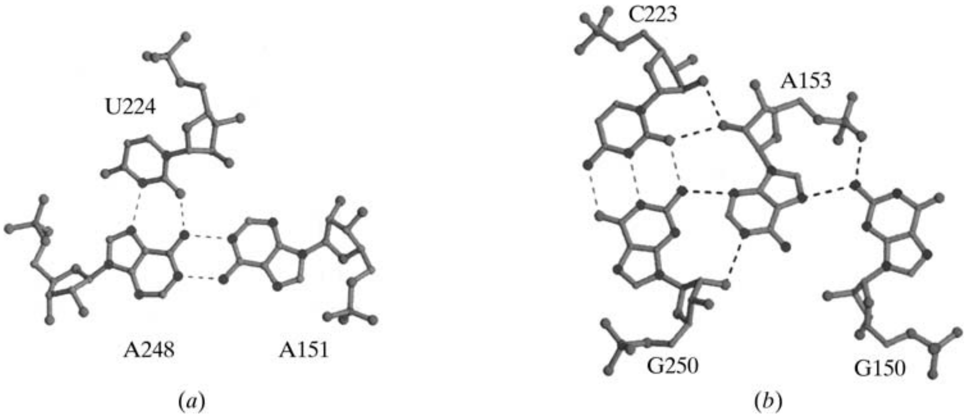


Figure 22.3.6.1
Hydrogen-bonding interactions in RNA tertiary structure. In (*a*), a triple base interaction is shown. In (*b*), G150 and A153 of a GAAA tetraloop participate in multiple hydrogen-bond interactions involving bases, riboses and phosphate. Reprinted with permission from Cate *et al.* (1996). Copyright (1996) American Association for the Advancement of Science.

GNRA tetraloop structure (N = any base, R = purine). This loop has a well defined structure, stabilized by hydrogen bonding and stacking involving its own bases, and it also presents further hydrogen-bonding groups that can dock into 'receptor' structures in other parts of the RNA molecule. This results in triple or quadruple base interactions (Fig. 22.3.6.1) that tie different parts of the RNA structure together; the parallel with hydrogen-bonding side chains in proteins is very strong. The 2′-hydroxyls of ribose groups are also used in some of these interactions (Fig. 22.3.6.1). Further ribose interactions involve interdigitated ribose groups that line the interfaces between adjacent helices such that pairs of riboses interact by hydrogen bonding through their 2′-hydroxyl groups, forming 'ribose zippers'. As many more RNA structures are determined experimentally, it is likely that more hydrogen-bonding motifs will be recognized, and their full role in RNA structure can be better assessed than at our present, imperfect state of knowledge.

22.3.7. Non-conventional hydrogen bonds

The vast majority of hydrogen bonds in biological macromolecules involve nitrogen and oxygen donors exclusively. Nevertheless, several other interactions have all the characteristics of hydrogen bonds and clearly contribute to structure and stability where they occur.

22.3.7.1. *C—H···O hydrogen bonds*

Sutor (1962) first summarized evidence for C—H···O hydrogen bonds following earlier suggestions by Pauling (1960), and current evidence has been nicely summarized in several articles (Derewenda *et al.*, 1995; Wahl & Sundaralingam, 1997; Jiang & Lai, 2002). The energy of C—H···O hydrogen bonds has been generally estimated as ~0.5 kcal mol^{-1} (about 10% of an N—H···O interaction) but may be higher, especially in hydrophobic environments. It also depends on the acidity of the C—H proton, with methylene (CH_2) and methyne (CH) groups being most favourable.

A number of examples of C—H···O hydrogen bonds can be found in nucleic acid structures (Wahl & Sundaralingam, 1997). The best known is that between the backbone O5′ oxygen and a purine C(8)—H or pyrimidine C(6)—H, when the bases are in the *anti* conformation. Another example is given by a U–U base pair, in which the two bases form a conventional N(3)—H···O(4) hydrogen bond and a C(5)—H···O hydrogen bond.

In proteins, two groups are regarded as being particularly significant (Derewenda *et al.*, 1995). These are the C^{ε}H of His side chains and the methylene H atoms of the main-chain α-carbon atoms. C—H···O hydrogen bonds involving His side chains have been found for the active-site His residues of proteins of the lipase/esterase family and in other proteins (Derewenda *et al.*, 1994). The C^{α}H atoms, however, appear to provide much more widespread C—H···O hydrogen bonding, especially in β-sheets, where they are directed towards the 'free' lone pairs of the main-chain C=O groups. C—H···O hydrogen bonds may thus play a previously unrecognized role in satisfying the hydrogen-bond potential of C=O groups. In general, Derewenda *et al.* (1995) find a significant number of C···O contacts that meet the criteria for C—H···O hydrogen bonds; the H···O distance peaks at 2.45 Å (C···O 3.5 Å), which is less than the van der Waals distance of 2.7 Å, and the angles indicate that the H atoms are directed at the acceptor lone-pair orbitals. C—H···O hydrogen bonds are also found in some cases to complete the tetrahedral coordination of internal water molecules (Desiraju & Steiner, 1999).

22.3.7.2. *Hydrogen bonds involving sulfur atoms*

Sulfur atoms are larger and have a more diffuse electron cloud than oxygen or nitrogen, but are nevertheless capable of participating in hydrogen bonds. Given that the radius of sulfur is ~0.4 Å greater than that of oxygen, hydrogen bonds can be assumed if the distance H···S is less than ~2.9 Å, or S···O(N) is less than ~3.9 Å, providing the angular geometry is right. In proteins, the SH group of cysteine can be a hydrogen-bond acceptor or donor, whereas the sulfur atoms in disulfide bonds and in Met side chains can act only as acceptors.

The clearest example of hydrogen bonding involving Cys residues is given by the NH···S hydrogen bonds first described in Fe-S proteins (Adman *et al.*, 1975); here, peptide NH groups are oriented to point directly at the S atoms of metal-bound Cys residues, with H···S distances of 2.4–2.9 Å. Similar NH···S hydrogen bonds are found in blue copper proteins, involving the Cys ligands. In these cases, the cysteine sulfur is deprotonated and therefore more negative, making it a stronger hydrogen-bond acceptor, and it is likely that hydrogen bonding to cysteine S^- atoms is common. A large survey of Cys and Met side chains in proteins has given evidence of both N—H···S and S—H···O hydrogen bonds involving the SH groups of Cys side chains (Gregoret *et al.*, 1991). In particular, Cys residues in helices frequently hydrogen bond to the main-chain C=O group four residues back in the helix in SH(n) ··· O($n-4$) interactions analogous to those seen for Ser and Thr residues in helices. On the other hand, O—H···S or N—H···S hydrogen bonds to the S atoms of Met or half-cystine side chains, although they do exist, are rare (Gregoret *et al.*, 1991; Ippolito *et al.*, 1990).

22.3.7.3. Amino–aromatic hydrogen bonding

Surveys of protein structures have shown that aromatic rings (of Trp, Tyr or Phe) are frequently in close association with side-chain NH groups of Lys, Arg, Asn, Gln or His (Burley & Petsko, 1986). Energy calculations further suggest that where an N—H group, as donor, is directed towards the centre of an aromatic ring, as acceptor, a hydrogen-bonded interaction with an energy of ~3 kcal mol^{-1} (about half that of a normal N—H···O or O—H···O hydrogen bond) can result (Levitt & Perutz, 1988). Whether the close associations observed by Burley & Petsko can truly be regarded as hydrogen bonds has been controversial, however. Mitchell *et al.* (1994) have analysed amino–aromatic interactions and shown that by far the most common form of association between sp^2 nitrogen atoms and aromatic rings involves approximately plane-to-plane stacking, which cannot represent hydrogen bonding. There are still, however, a significant number of cases where the H atoms of N—H groups are directed towards aromatic rings, and these represent genuine hydrogen bonds (Mitchell *et al.*, 1994). It is clearly essential to consider the donor–acceptor geometry, both distances and angles, before assuming an amino–aromatic hydrogen bond; the N···ring distance should be less than ~3.8 Å, and N—H··· C angle greater than 120°, where C is the ring centre (Mitchell *et al.*, 1994).

The author gratefully acknowledges many useful discussions with Guy Dodson and Rod Hubbard (University of York), and many other colleagues. Clyde Smith is thanked for help with figures, and research support from the Health Research Council of New Zealand and the Tertiary Education Commission of New Zealand through its support of the Maurice Wilkins Centre for Molecular Biodiscovery is acknowledged.

References

Adman, E., Watenpaugh, K. D. & Jensen, L. H. (1975). *N—H···S hydrogen bonds in Peptococcus aerogenes ferredoxin, Clostridium pasteurianum rubredoxin and Chromatium high potential iron protein. Proc. Natl Acad. Sci. USA*, **72**, 4854–4858.

Alber, T., Dao-pin, S., Wilson, K., Wozniak, J. A., Cook, S. P. & Matthews, B. W. (1987). *Contributions of hydrogen bonds of Thr 157 to the thermodynamic stability of phage T4 lysozyme. Nature (London)*, **330**, 41–46.

Artymiuk, P. J. & Blake, C. C. F. (1981). *Refinement of human lysozyme at 1.5 Å resolution. Analysis of non-bonded and hydrogen-bonded interactions. J. Mol. Biol.* **152**, 737–762.

Baker, E. N. (1995). *Solvent interactions with proteins as revealed by X-ray crystallographic studies.* In *Protein–Solvent Interactions*, edited by R. B. Gregory, pp. 143–189. New York: Marcel Dekker Inc.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Blundell, T., Barlow, D., Borkakoti, N. & Thornton, J. (1983). *Solvent-induced distortions and the curvature of α-helices. Nature (London)*, **306**, 281–283.

Bordo, D. & Argos, P. (1994). *The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. J. Mol. Biol.* **243**, 504–519.

Burley, S. K. & Petsko, G. A. (1986). *Amino–aromatic interactions in proteins. FEBS Lett.* **203**, 139–143.

Cate, J. H., Gooding, A. R., Podell, E., Zhou, K., Golden, B. L., Kundrot, C. E., Cech, T. R. & Doudna, J. A. (1996). *Crystal structure of a group I ribozyme domain: principles of RNA packing. Science*, **273**, 1678–1685.

Cleland, W. W., Frey, P. A. & Gerlt, J. A. (1998). *The low barrier hydrogen bond in enzymatic catalysis. J. Biol. Chem.* **273**, 25529–25532.

Derewenda, Z. S., Derewenda, U. & Kobos, P. (1994). *(His)Cε—H···O═C< hydrogen bond in the active site of serine hydrolases. J. Mol. Biol.* **241**, 83–93.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins. J. Mol. Biol.* **252**, 248–262.

Desiraju, G. R. (2011). *Hydrogen bonds: a bond by any other name. Angew. Chem. Int. Ed.* **50**, 52–59.

Desiraju, G. R. & Steiner, T. (1999). *The Weak Hydrogen Bond in Structural Chemistry and Biology*. Oxford University Press.

Edwards, R. A., Baker, H. M., Whittaker, M. M., Whittaker, J. W. & Baker, E. N. (1998). *Crystal structure of Escherichia coli manganese superoxide dismutase at 2.1 Å resolution. J. Biol. Inorg. Chem.* **3**, 161–171.

Eswar, N. & Ramakrishnan, C. (2000). *Deterministic features of side-chain main-chain hydrogen bonds in globular protein structures. Protein Eng.* **13**, 227–238.

Fersht, A. R. & Serrano, L. (1993). *Principles in protein stability derived from protein engineering experiments. Curr. Opin. Struct. Biol.* **3**, 75–83.

Fersht, A. R., Shi, J.-P., Knill-Jones, J., Lowe, D. M., Wilkinson, A. J., Blow, D. M., Brick, P., Carter, P., Waye, M. M. Y. & Winter, G. (1985). *Hydrogen bonding and biological specificity analysed by protein engineering. Nature (London)*, **314**, 235–238.

Flocco, M. M. & Mowbray, S. L. (1995). *Strange bedfellows: interactions between acidic side-chains in proteins. J. Mol. Biol.* **254**, 96–105.

Gao, J., Bosco, D. A., Powers, E. T. & Kelly, J. W. (2009). *Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. Nat. Struct. Mol. Biol.* **16**, 684–690.

Gregoret, L. M., Rader, S. D., Fletterick, R. J. & Cohen, F. E. (1991). *Hydrogen bonds involving sulfur atoms in proteins. Proteins*, **9**, 99–107.

Gutteridge, A. & Thornton, J. M. (2005). *Understanding Nature's catalytic toolkit. Trends Biochem. Sci.* **30**, 622–629.

Hagler, A. T., Huler, E. & Lifson, S. (1974). *Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Am. Chem. Soc.* **96**, 5319–5327.

Harper, E. T. & Rose, G. D. (1993). *Helix stop signals in proteins and peptides: the capping box. Biochemistry*, **32**, 7605–7609.

Huggins, M. L. (1971). *50 years of hydrogen bonding theory. Angew. Chem. Int. Ed. Engl.* **10**, 147–208.

Ippolito, J. A., Alexander, R. S. & Christianson, D. W. (1990). *Hydrogen bond stereochemistry in protein structure and function. J. Mol. Biol.* **215**, 457–471.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures*. New York: Springer-Verlag.

Jiang, L. & Lai, L. (2002). *CH···O hydrogen bonds at protein–protein interfaces. J. Biol. Chem.* **277**, 37732–37740.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–64.

Langkilde, A., Kristensen, S. M., Lo Leggio, L., Molgaard, A., Jensen, J. H., Houk, A. R., Navarro Poulsen, J.-C., Kaupinnen, S. & Larsen, S. (2008). *Short strong hydrogen bonds in proteins: a case study of rhamnogalacturonan acetylesterase. Acta Cryst.* D**64**, 851–863.

Legon, A. C. & Millen, D. J. (1987). *Directional character, strength, and nature of the hydrogen bond in gas-phase dimers. Acc. Chem. Res.* **20**, 39–45.

Levitt, M. & Perutz, M. F. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

Liu, Z., Wang, G., Li, Z. & Wang, R. (2008). *Geometrical preferences of the hydrogen bonds in protein-ligand interfaces derived from statistical surveys and quantum mechanics calculations. J. Chem. Theory Comput.* **4**, 1959–1973.

McDonald, I. K. & Thornton, J. M. (1994*a*). *Satisfying hydrogen bonding potential in proteins. J. Mol. Biol.* **238**, 777–793.

McDonald, I. K. & Thornton, J. M. (1994*b*). *The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng.* **8**, 217–224.

Matthews, B. W. (1972). *The γ turn. Evidence for a new folded conformation in proteins. Macromolecules*, **5**, 818–819.

Mitchell, J. B. O., Nandi, C. L., McDonald, I. K., Thornton, J. M. & Price, S. L. (1994). *Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? J. Mol. Biol.* **239**, 315–331.

Nemethy, G. & Printz, M. P. (1972). *The γ turn, a possible folded conformation of the polypeptide chain. Comparison with the β turn. Macromolecules*, **5**, 755–758.

Pace, C. N. (2009). *Energetics of protein hydrogen bonds. Nat. Struct. Mol. Biol.* **16**, 681–682.

Pauling, L. (1960). *The Nature of the Chemical Bond*, 3rd ed. Ithaca: Cornell University Press.

Pauling, L. & Corey, R. B. (1951). *Configurations of polypeptide chains with favoured orientations around single bonds: two new pleated sheets. Proc. Natl Acad. Sci. USA*, **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl Acad. Sci. USA*, **37**, 205–211.

Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). *Three-dimensional structure of a hammerhead ribozyme. Nature (London)*, **372**, 68–74.

Presta, L. G. & Rose, G. D. (1988). *Helix signals in proteins. Science*, **240**, 1632–1641.

Richardson, J. S., Arendall, W. B. & Richardson, D. C. (2003). *New tools and data for improving structures, using all-atom contacts*. *Methods Enzymol.* **374**, 385–412.

Richardson, J. S., Getzoff, E. D. & Richardson, D. C. (1978). *The β-bulge: a common small unit of nonrepetitive protein structure. Proc. Natl Acad. Sci. USA*, **75**, 2574–2578.

Richardson, J. S. & Richardson, D. C. (1988). *Amino acid preferences for specific locations at the ends of α-helices. Science*, **240**, 1648–1652.

Savage, H. J., Elliott, C. J., Freeman, C. M. & Finney, J. L. (1993). *Lost hydrogen bonds and buried surface area: rationalising stability in globular proteins. J. Chem. Soc. Faraday Trans.* **89**, 2609–2617.

Schellman, C. (1980). *The alpha-L conformation at the ends of helices.* In *Protein Folding*, edited by R. Jaenicke, pp. 53–61. Amsterdam: Elsevier.

Schutz, C. N. & Warshel, A. (2004). *The low barrier hydrogen bond (LBHB) proposal revisited: the case of the Asp···His pair in serine proteases. Proteins Struct. Funct. Bioinform.* **55**, 711–723.

Stickle, D. F., Presta, L. G., Dill, K. A. & Rose, G. D. (1992). *Hydrogen bonding in globular proteins. J. Mol. Biol.* **226**, 1143–1159.

Sutor, D. J. (1962). *The C—H···O hydrogen bond in crystals. Nature (London)*, **195**, 68–69.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H···O═C hydrogen bond. 1. Lone pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1988). *Distribution of water around amino acids in proteins. J. Mol. Biol.* **202**, 637–657.

Thanki, N., Umrania, Y., Thornton, J. M. & Goodfellow, J. M. (1991). *Analysis of protein main-chain solvation as a function of secondary structure. J. Mol. Biol.* **221**, 669–691.

Torrance, J. W., Holliday, G. L., Mitchell, J. B. O. & Thornton, J. M. (2007). *The geometry of interactions between catalytic residues and their substrates*. *J. Mol. Biol.* **369**, 1140–1152.

Wahl, M. C. & Sundaralingam, M. (1997). *C—H···O hydrogen bonding in biology*. *Trends Biochem. Sci.* **22**, 97–102.

Weaver, T. M. (2000). *The π-helix translates structure into function*. *Protein Sci.* **9**, 201–206.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins. Protein Sci.* **3**, 1224–1235.

International Tables for Crystallography (2012). Vol. F, Chapter 22.4, pp. 730–735.

Chapter 22.4. Electrostatic interactions in proteins

K. A. Sharp

22.4.1. Introduction

Electrostatic interactions play a key role in determining the structure, stability, binding affinity, chemical properties, and hence the biological reactivity, of proteins and nucleic acids. Interactions where electrostatics play an important role include:

(1) Ligand/substrate association. Long-range electrostatic forces can considerably enhance association rates by facilitating translational and rotational diffusion or by reduction in the dimensionality of the diffusion space.

(2) Binding affinity. Tight specific binding is often a prerequisite for biological activity, and electrostatics make important contributions to desolvation and formation of chemically complementary interactions during binding.

(3) Modification of chemical and physical properties of functional groups such as cofactors (haems, metal ions *etc.*), alteration of the ionization energy pK_a of side chains and shifting of redox midpoints.

(4) The creation of potentials or fields in the active sites to stabilize functionally important charged or dipolar intermediates in processes such as catalysis.

In this chapter I will discuss, within the framework of classical electrostatics, how such effects can be modelled starting from the structural information provided by X-ray crystallography. Nevertheless, many of the concepts of classical electrostatics can be used in combination with molecular dynamics (MD), quantum mechanics (QM) and other computational methods to study a wider range of macromolecular properties, for example specific protein motions, the breaking or forming of bonds, determination of intrinsic pK_a's, determination of electronic energy levels *etc.*

The central aim in studying the electrostatic properties of macromolecules is to take the structural information provided by crystallography (typically the atomic coordinates, although *B*-factor information may also be of use) and obtain a realistic description of the electrostatic potential distribution $\varphi(\mathbf{r})$. The electrostatic potential distribution can then be used in a variety of ways: (i) graphical analysis may reveal deeper aspects of the structure and help identify functionally important regions or active sites; (ii) the potentials may be used to calculate energies and forces, which can then be used to calculate equilibrium or kinetic properties; and (iii) the electrostatic potentials may be used in conjunction with other computational methods such as QM and MD.

Three problems must be solved to obtain the electrostatic potential distribution. The first is to model the macromolecular charge distribution, usually by specifying the location and charge of all its atoms. Although the coordinates of the molecule are determined by crystallographic methods, the charge distribution is not. A number of atomic charge distributions have been developed for proteins and nucleic acids using quantum mechanical methods and/or parameterization to different experimental data. The second problem is that the positions of the water molecules and solvent ions are generally not known. (Water molecules and ions seen in even the best crystal structures usually constitute a small fraction of the total important in solvating the molecule. Moreover, the orientation of the crystallographic water molecules, crucial in determining the electrostatic potential, is rarely known.) Within the framework of classical electrostatics, inclusion of the *effect* of the solvating water molecules and ions is handled not by treating them explicitly, but implicitly in terms of an 'electrostatic response' to the field created by the molecular charge distribution. The third problem is that incorporation of the available structural information at atomic resolution results in a complicated spatial distribution of charge, dielectric response *etc.* Numerical methods for rapidly and accurately solving the electrostatic equations that determine the potential are therefore essential.

22.4.2. Theory

22.4.2.1. The response of the system to electrostatic fields

The response to the electrostatic field arising from the molecular charge distribution arises from three physical processes: electronic polarization, reorientation of permanent dipolar groups and redistribution of mobile ions in the solvent. Movement of ionized side chains, if significant, is sometimes viewed as part of the dielectric response of the protein, and sometimes explicitly as a conformational change of the molecule.

Electronic polarizability can be represented either by point inducible dipoles (Warshel & Åqvist, 1991) or by a dielectric constant. The latter approach relates the electrostatic polarization, $\mathbf{P}(\mathbf{r})$ (the mean dipole moment induced in some small volume V) to the Maxwell (total) field, $\mathbf{E}(\mathbf{r})$, and the local dielectric constant representing the response of that volume, $\varepsilon(\mathbf{r})$, according to

$$\mathbf{P}(\mathbf{r}) = [\varepsilon(\mathbf{r}) - 1]\mathbf{E}(\mathbf{r})/4\pi. \qquad (22.4.2.1)$$

The contribution of electronic polarizability to the dielectric constant of most organic material and water is fairly similar. It can be evaluated by high-frequency dielectric measurements or the refractive index, and it is in the range 2–2.5.

The reorientation of groups such as the peptide bond or surrounding water molecules which have large permanent dipoles is an important part of the overall response. This response too may be treated using a dielectric constant, *i.e.* using equation (22.4.2.1) with a larger value of the dielectric constant that incorporates the additional polarization from dipole reorientation. An alternative approach to equation (22.4.2.1) for treating the dipole reorientation contribution of water surrounding the macromolecules is the Langevin dipole model (Lee *et al.*, 1993; Warshel & Åqvist, 1991; Warshel & Russell, 1984). Four factors determine the degree of response from permanent dipoles: (i) the dipole-moment magnitude; (ii) the density of such groups in the protein or solvent; (iii) the freedom of such groups to reorient; and (iv) the degree of cooperativity between dipole motions. Thus, water has a high dielectric constant (ε = 78.6 at 25 °C). For electrostatic models based on dielectric theory, the experimental solvent dielectric constant, reflecting the contribution of electronic polarizability and dipole reorientation, is usually used. From consideration of the four factors that determine the

Copyright © 2012 International Union of Crystallography

dielectric response, macromolecules would be expected to have a much lower dielectric constant than the solvent. Indeed, theoretical studies of the dielectric behaviour of amorphous protein solids (Gilson & Honig, 1986; Nakamura *et al.*, 1988) and the interior of proteins in solution (Simonson & Brooks, 1996; Simonson & Perahia, 1995; Smith *et al.*, 1993), and experimental measurements (Takashima & Schwan, 1965) provide an estimate of $\varepsilon = 2.5$–4 for the contribution of dipolar groups to the protein dielectric.

The Langevin model can account for the saturation of the response at high fields that occurs if the dipoles become highly aligned with the field. The dielectric model can also be extended to incorporate saturation effects (Warwicker, 1994), although there is a compensating effect of electrostriction, which increases the local dipole density (Jayaram, Fine *et al.*, 1989). While the importance of saturation effects would vary from case to case, linear solvent dielectric models have proven sufficiently accurate for most protein applications to date.

Charge groups on molecules will attract solvent counter-ions and repel co-ions. The most common way of treating this charge rearrangement is *via* the Boltzmann model, where the net charge density of mobile ions is given by

$$\rho^m(\mathbf{r}) = \sum_i z_i e c_i^o \exp[-z_i e \varphi(\mathbf{r})/kT], \qquad (22.4.2.2)$$

where c_i^o is the bulk concentration of an ion of type i, valence z_i, and $\varphi(\mathbf{r})$ is the average potential (an approximation to the potential of mean force) at position $\mathbf{r}$. The Boltzmann approach neglects the effect of ion size and correlations between ion positions. Other models for the mobile-ion behaviour that account for these effects are integral equation models and MC models (Bacquet & Rossky, 1984; Murthy *et al.*, 1985; Olmsted *et al.*, 1989, 1991; Record *et al.*, 1990). These studies show that ion size and correlation effects do not compromise the Boltzmann model significantly for monovalent (1–1) salts at mid-range concentrations 0.001–0.5 *M*, and consequently it is widely used for modelling salt effects in proteins and nucleic acids.

22.4.2.2. Dependence of the potential on the charge distribution

The potential at a point in space, $\mathbf{r}$, arising from some charge density distribution $\rho(\mathbf{s})$ and some dipole density distribution $\mathbf{P}(\mathbf{s})$ (which includes polarization) is given by

$$\varphi(\mathbf{r}) = \int \rho(\mathbf{s})/(|\mathbf{s}-\mathbf{r}|) + \mathbf{P}(\mathbf{s})(\mathbf{s}-\mathbf{r})/(|\mathbf{s}-\mathbf{r}|^3)\,\mathrm{d}\mathbf{s}. \qquad (22.4.2.3)$$

The total charge distribution is the sum of the explicit charge distribution on the molecule and that from the mobile solvent ion distribution, $\rho = \rho^e + \rho^m$. Substituting for the dielectric polarization using equation (22.4.2.1) and for the mobile ion charge distribution using equation (22.4.2.2), the potential may be expressed in terms of a partial differential equation, the Poisson–Boltzmann (PB) equation:

$$\nabla\varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r}) + 4\pi\sum_i z_i e c_i^o \exp[-z_i e\varphi(\mathbf{r})/kT] + 4\pi\rho^e(\mathbf{r}) = 0, \qquad (22.4.2.4)$$

which relates the potential, molecular charge and dielectric distributions, $\varphi(\mathbf{r})$, $\rho^e(\mathbf{r})$ and $\varepsilon(\mathbf{r})$, respectively. Contributions to the polarizability from electrons, a molecule's permanent dipoles and solvent dipoles are incorporated into this model by using an appropriate value for the dielectric for each region of protein and solvent. Values for protein atomic charges, radii and dielectric constants suitable for use with the Poisson–Boltzmann equation are available in the literature (Jean-Charles *et al.*, 1990; Mohan *et al.*, 1992; Simonson & Brünger, 1994; Sitkoff *et al.*, 1994). For protein applications, the Boltzmann term in equation (22.4.2.4) is usually linearized to become $-8\pi\varphi(\mathbf{r})I/kT$ where I is the ionic strength, whereas for nucleic acids and molecules of similarly high charge density the full nonlinear equation is used.

22.4.2.3. The concepts of screening, reaction potentials, solvation, dielectric, polarity and polarizability

Application of a classical electrostatic view to macromolecular electrostatics involves a number of useful concepts that describe the physical behaviour. It should first be recognized that the potential at a particular charged atom i includes three physically distinct contributions. The first is the direct or Coulombic potential of j at i. The second is the potential at i generated by the polarization (of a molecule, water and ion atmosphere) induced by j. This is often referred to as the screening potential, since it opposes the direct Coulombic potential. The third arises from the polarization induced by i itself. This is often referred to as the reaction or self-potential, or if solvent is involved, as the solvation potential.

When using models that apply the concept of a dielectric constant (a measure of polarizability) to a macromolecule, it is important to distinguish between polarity and polarizability. Briefly, polarity may be thought of as describing the density of charged and dipolar groups in a particular region. Polarizability, by contrast, refers to the *potential* for reorganizing charges, orienting dipoles and inducing dipoles. Thus polarizability depends both on the polarity and the freedom of dipoles to reorganize in response to an applied electric field. When a protein is folding or undergoing a large conformational rearrangement, the peptide groups may be quite free to reorient. In the folded protein, these may become spatially organized so as to stabilize another charge or dipole, creating a region with high polarity, but with low polarizability, since there is much less ability to reorient the dipolar groups in response to a new charge or dipole without significant disruption of the structure. Thus, while there is still some discussion about the value and applicability of a protein dielectric constant, it is generally agreed that the interior of a macromolecule is a less polarizable environment compared to solvent. This difference in polarizability has a significant effect on the potential distribution.

Formally charged groups on proteins, particularly the longer side chains on the surface of proteins, Arg, Lys, and to a lesser extent Glu and Asp, have the ability to alter their conformation in response to electrostatic fields. In addition, information about fluctuations about their mean position may need to be included in calculating average properties. Three approaches to modelling protein formal charge movements can be taken. The first is to treat the motions within the dielectric response. In this approach, the protein may be viewed as having a dielectric higher than 2.5–4 in the regions of these charged groups, particularly at the surface, where the concentration and mobility of these groups may give an effective dielectric of 20 or more (Antosiewicz *et al.*, 1994; Simonson & Perahia, 1995; Smith *et al.*, 1993). A second approach is to model the effect of charge motions on the electrostatic quantity of interest explicitly, *e.g.* with MD simulations (Langsetmo *et al.*, 1991; Wendoloski & Matthew, 1989). This involves generating an ensemble of structures with different atomic charge distributions. The third approach is based on the fact that one is often interested in a specific biological process $A \rightarrow B$ in which one can evaluate the structure of the protein in states A and B (experimentally or by modelling), and any change in average charge positions is incorporated at the level of different average explicit charge distribution inputs for the calculation, modelling only the electronic, dipolar and salt contributions as the response.

The term 'effective' dielectric constant is sometimes used in the literature to describe the strength of interaction between two charges, q_1 and q_2. This is defined as the ratio of the observed or calculated interaction strength, U, to that expected between the same two charges in a vacuum:

$$\varepsilon^{\rm eff} = [(q_1 q_2)/r_{12}]/U, \qquad (22.4.2.5)$$

where r_{12} is the distance between the charges. If the system were completely homogeneous in terms of its electrostatic response and involved no charge rearrangement then $\varepsilon^{\rm eff}$ would describe the dielectric constant of the medium containing the charges. This is generally never the case: the strength of interaction in a protein system is determined by the net contribution from protein, solvent and ions, so $\varepsilon^{\rm eff}$ does not give information about the dielectric property of any particular region of space. In fact, in the same system different charge–charge interactions will generally yield different values of $\varepsilon^{\rm eff}$. Thus $\varepsilon^{\rm eff}$ is really no more than its definition – a measure of the strength of interaction – and it cannot be used directly to answer questions about the protein dielectric constant, for example. Rather, it is one of the quantities that one aims to extract from theoretical models to compare with an experiment.

22.4.2.4. Calculation of energies and forces

Once the electrostatic potential distribution has been obtained, calculation of experimental properties usually requires evaluation of the electrostatic energy or force. For a linear system (where the dielectric and ionic responses are linear) the electrostatic free energy is given by

$$\Delta G^{\rm el} = 1/2 \sum_i \varphi_i q_i, \qquad (22.4.2.6)$$

where $\varphi_{\rm i}$ is the potential at an atom with charge q_i. The most common source of nonlinearity is the Boltzmann term in the PB equation (22.4.2.4) for highly charged molecules such as nucleic acids. The total electrostatic energy in this case is (Reiner & Radke, 1990; Sharp & Honig, 1990; Zhou, 1994)

$$\Delta G^{\rm el} = \int_V \{\rho^e \varphi - (\varepsilon E^2/8\pi) - kT \sum_i c_i^0 [\exp(-z_i e\varphi/kT) - 1]\}\, {\rm d}\mathbf{r}, \qquad (22.4.2.7)$$

where the integration is now over all space.

The general expression for the electrostatic force on a charge q is given by the gradient of the total free energy with respect to that charge's position,

$$\mathbf{f}_q = -\nabla_{\mathbf{r}q}(G^{\rm el}). \qquad (22.4.2.8)$$

If the movement of that charge does not affect the potential distribution due to the other charges and dipoles, then equation (22.4.2.8) can be evaluated using the 'test charge' approach, in which case the force depends only on the gradient of the potential or the field at the charge:

$$\mathbf{f} = q\mathbf{E}. \qquad (22.4.2.9)$$

However, in a system like a macromolecule in water, which has a non-homogeneous dielectric, forces arise between a charge and any dielectric boundary due to image charge (reaction potential) effects. A similar effect to the 'dielectric pressure' force arises from solvent-ion pressure at the solute–solvent boundary. This results in a force acting to increase the solvent exposure of charged and polar atoms. An expression for the force that includes these effects has been derived within the PB model (Gilson *et al.*, 1993):

$$\mathbf{f} = \rho^e \mathbf{E} - (1/2)E^2 \nabla\varepsilon - kT \sum_i c_i^0 [\exp(-z_i e\varphi/kT) - 1]\nabla A, \qquad (22.4.2.10)$$

where A is a function describing the accessibility to solvent ions, which is 0 inside the protein, and 1 in the solvent, and whose gradient is nonzero only at the solute–solvent surface. Similarly, in a two-dielectric model (solvent plus molecule) the gradient of ε is nonzero only at the molecular surface. The first term accounts for the force acting on a charge due to a field, as in equation (22.4.2.9), while the second and third terms account for the dielectric surface pressure and ionic atmosphere pressure terms respectively. Equation (22.4.2.10) has been used to combine the PB equation and molecular mechanics (Gilson *et al.*, 1995).

22.4.2.5. Numerical methods

A variety of numerical methods exist for calculating electrostatic potentials of macromolecules. These include numerical solution of self-consistent field electrostatic equations, which has been used in conjunction with the protein dipole–Langevin dipole method (Lee *et al.*, 1993). Numerical solution of the Poisson–Boltzmann equation requires the solution of a three-dimensional partial differential equation, which can be nonlinear. Many numerical techniques, some developed in engineering fields to solve differential equations, have been applied to the PB equation. These include finite-difference methods (Bruccoleri *et al.*, 1996; Gilson *et al.*, 1988; Nicholls & Honig, 1991; Warwicker & Watson, 1982), finite-element methods (Rashin, 1990; Yoon & Lenhoff, 1992; Zauhar & Morgan, 1985), multigridding (Holst & Saied, 1993; Oberoi & Allewell, 1993), conjugate-gradient methods (Davis & McCammon, 1989) and fast multipole methods (Bharadwaj *et al.*, 1994; Davis, 1994). Methods for treating the nonlinear PB equation include under-relaxation (Jayaram, Sharp & Honig, 1989) and powerful inexact Newton methods (Holst *et al.*, 1994). The nonlinear PB equation can also be solved *via* a self-consistent field approach, in which one calculates the potential using equation (22.4.2.5), then the mobile charge density is calculated using equation (22.4.2.3), and the procedure is repeated until convergence is reached (Pack & Klein, 1984; Pack *et al.*, 1986). The method allows one to include more elaborate models for the ion distribution, for example incorporating the finite size of the ions (Pack *et al.*, 1993). Approximate methods based on spherical approximations (Born-type models) have also been used (Schaeffer & Frommel, 1990; Still *et al.*, 1990). Considerable numerical progress has been made in finite methods, and accurate rapid algorithms are available. The reader is referred to the original references for numerical details.

22.4.3. Applications

An exhaustive list of applications of classical electrostatic modelling to macromolecules is beyond the scope of this chapter. Three general areas of application are discussed.

22.4.3.1. Electrostatic potential distributions

Graphical analysis of electrostatic potential distributions often reveals features about the structure that complement analysis of the atomic coordinates. For example, Fig. 22.4.3.1(*a*) shows the

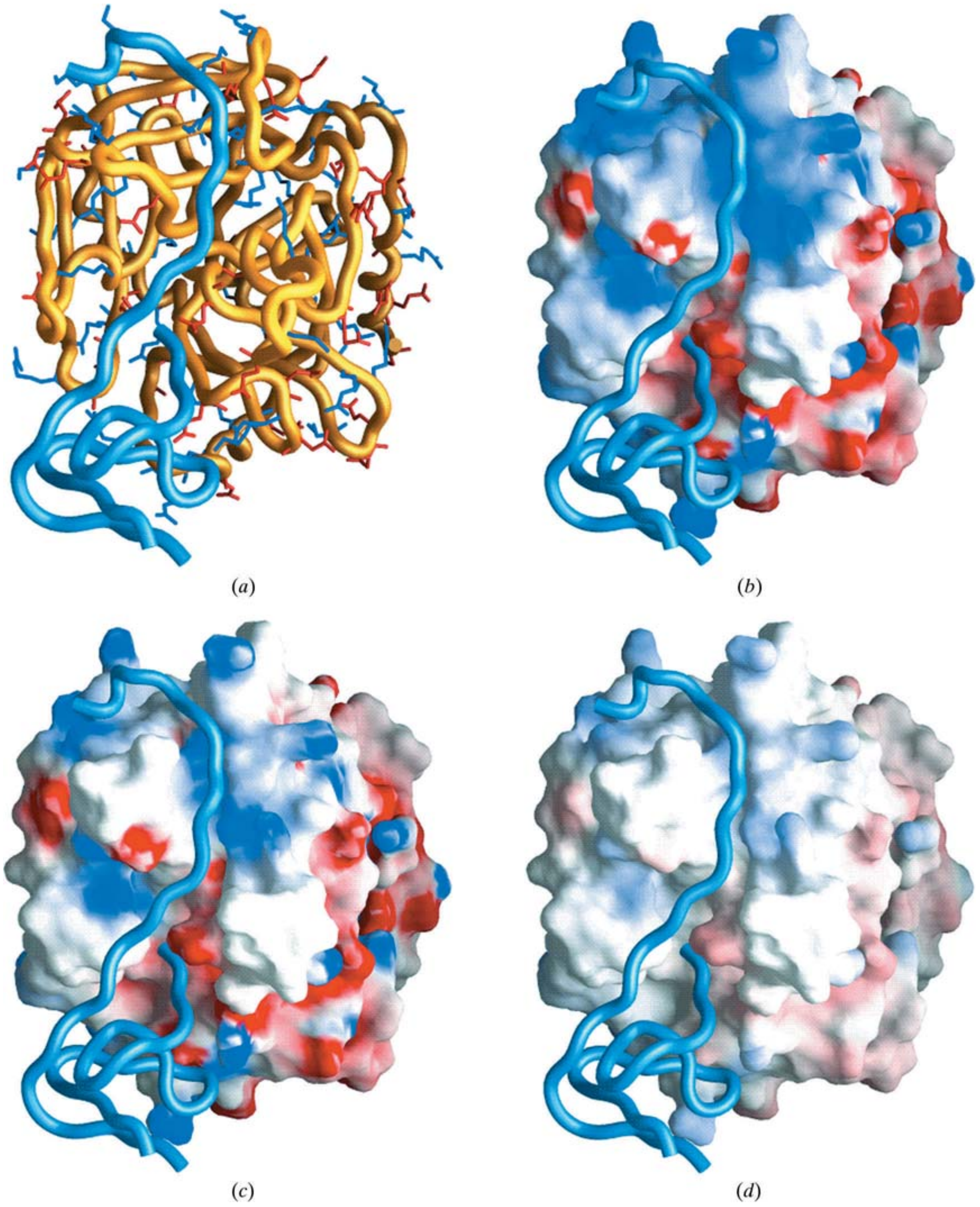


Figure 22.4.3.1
(*a*) The proteolytic enzyme thrombin (yellow backbone worm) complexed with an inhibitor, hirudin (blue backbone worm). The negatively charged (red) and positively charged (blue) side chains of thrombin are shown in bond representation. (*b*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated at zero ionic strength. (*c*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated at physiological ionic strength (0.145 *M*). (*d*) Solvent-accessible surface of thrombin coded by electrostatic potential (blue: positive, red: negative). Hirudin is shown as a blue backbone worm. Potential is calculated using the same polarizability for protein and solvent.

distribution of charged residues in the binding site of the proteolytic enzyme thrombin. Fig. 22.4.3.1(*b*) shows the resulting electrostatic potential distribution on the protein surface. The basic (positive) region in the fibrinogen binding site, which could be inferred from close inspection of the distribution of charged residues in Fig. 22.4.3.1(*a*), is clearly more apparent in the potential distribution. Fig. 22.4.3.1(*c*) shows the effect of increasing ionic strength on the potential distribution, shrinking the regions of strong potential. Fig. 22.4.3.1(*d*) is calculated assuming the same dielectric for the solvent and protein. The more uniform potential distribution compared to Fig. 22.4.3.1(*b*) shows the focusing effect that the low dielectric interior has on the field emanating from charges in active sites and other cleft regions.

22.4.3.2. Charge-transfer equilibria

Charge-transfer processes are important in protein catalysis, binding, conformational changes and many other functions. The primary examples are acid–base equilibria, electron transfer and ion binding, in which the transferred species is a proton, an electron or a salt ion, respectively. The theory of the dependence of these three equilibria within the classical electrostatic framework can be treated in an identical manner, and will be illustrated with acid–base equilibria. A titratable group will have an intrinsic ionization equilibrium, expressed in terms of a known intrinsic pK_a^0, where $pK_a^0 = -\log_{10}(K_a^0)$, K_a^0 is the dissociation constant for the reaction $\mathrm{H}^+A = \mathrm{H}^+ + A$ and A can be an acid or a base. The pK_a^0 is determined by all the quantum-chemical, electrostatic and environmental effects operating on that group in some reference state. For example, a reference state for the aspartic acid side-chain ionization might be the isolated amino acid in water, for which $pK_a^0 = 3.85$. In the environment of the protein, the pK_a will be altered by three electrostatic effects. The first occurs because the group is positioned in a protein environment with a different polarizability, the second is due to interaction with permanent dipoles in the protein, the third is due to charged, perhaps titratable, groups. The effective pK_a is given by

$$pK_a = pK_a^0 + (\Delta\Delta G^{\mathrm{rf}} + \Delta\Delta G^{\mathrm{perm}} + \Delta\Delta G^{\mathrm{tit}})/2.303kT, \tag{22.4.3.1}$$

where the factor of $1/2.303kT$ converts units of energy to units of pK_a. The first contribution, $\Delta\Delta G^{\mathrm{rf}}$, arises because the completely solvated group induces a strong favourable reaction field (see Section 22.4.2.3) in the high dielectric water, which stabilizes the charged form of the group. (The neutral form is also stabilized by the solvent reaction field induced by any dipolar groups, but to a lesser extent.) Desolvating the group to any degree by moving it into a less polarizable environment will preferentially destabilize the charged form of that group, shifting the pK_a by an amount

$$\Delta\Delta G^{\mathrm{rf}} = (1/2)\sum_i \left(q_i^d \Delta\varphi_i^{\mathrm{rf},\,d} - q_i^p \Delta\varphi_i^{\mathrm{rf},\,p}\right), \tag{22.4.3.2}$$

where q_i^p and q_i^d are the charge distributions on the group, $\Delta\varphi_i^{\mathrm{rf},\,p}$ and $\Delta\varphi_i^{\mathrm{rf},\,d}$ are the changes in the group's reaction potential upon moving it from its reference state into the protein, in the protonated (superscript p) and deprotonated (superscript d) forms, respectively, and the sum is over the group's charges. The contribution of the permanent dipoles is given by

$$\Delta\Delta G^{\mathrm{tit}} = \sum_i \left(q_i^d - q_i^p\right)\varphi_i^{\mathrm{perm}}, \tag{22.4.3.3}$$

where $\varphi_i^{\mathrm{perm}}$ is the interaction potential at the ith charge due to all the permanent dipoles in the protein, including the effect of screening. It is observed that intrinsic pK_a's of groups in proteins are rarely shifted by more than 1 pK_a unit, indicating that the effects of desolvation are often compensated to a large degree by the $\Delta\Delta G^{\mathrm{perm}}$ term (Antosiewicz *et al.*, 1994). The final term accounts for the contribution of all the other charged groups:

$$\Delta\Delta G^{\mathrm{tit}} = \sum_i \left(q_i^d \langle\varphi_i\rangle^d_{\mathrm{pH},\,c,\,\Delta V} - q_i^p \langle\varphi_i\rangle^p_{\mathrm{pH},\,c,\,\Delta V}\right), \tag{22.4.3.4}$$

where $\langle\varphi_i\rangle$ is the mean potential at group charge i from all the other titratable groups. The charge states of the other groups in the protein depend in turn on their intrinsic 'pK_a's', on the external pH if they are acid–base groups, the external redox potential, ΔV, if they are redox groups and the concentration of ions, c, if they are ion-binding sites, as indicated by the subscript to $\langle\varphi_i\rangle$. Moreover, the charge state of the group itself will affect the equilibrium at the other sites. Because of this linkage, exact determination of the complete charged state of a protein is a complex procedure. If there are N such groups, the rigorous approach is to compute the titration-state partition function by evaluating the relative electrostatic free energies of all 2^N ionization states for a given set of pH, c, ΔV. From this one may calculate the mean ionization state of any group as a function of pH, ΔV *etc*. For large N this becomes impractical, but various approximate schemes work well, including a Monte Carlo procedure (Beroza *et al.*, 1991; Yang *et al.*, 1993) or partial evaluation of the titration partition function by clustering the groups into strongly interacting sub-domains (Bashford & Karplus, 1990; Gilson, 1993; Yang *et al.*, 1993).

Calculation of ion-binding and electron-transfer equilibria in proteins proceeds exactly as for calculation of acid–base equilibria, the results usually being expressed in terms of an association constant, K_a, or a redox midpoint potential E_m (defined as the external reducing potential at which the group is half oxidized and half reduced, usually at pH 7), respectively.

22.4.3.3. Electrostatic contributions to binding energy

The electrostatic contribution to the binding energy of two molecules is obtained by taking the difference in total electrostatic energies in the bound (AB) and unbound $A + B$ states. For the linear case,

$$\Delta\Delta G^{\mathrm{elec}}_{\mathrm{bind}} = (1/2)\sum_i^{N_A} q_i^A(\varphi_i^{AB} - \varphi_i^A) + (1/2)\sum_j^{N_B} q_i^B(\varphi_j^{AB} - \varphi_j^B), \tag{22.4.3.5}$$

where the first and second sums are over all charges in molecule A and B, respectively, and φ^x is the total potential produced by $x = A$, B, or AB. From equation (22.4.3.5), it should be noted that the electrostatic free energy change of each molecule has contributions from intermolecular charge–charge interactions, and from changes in the solvent reaction potential of the molecule itself when solvent is displaced by the other molecule. Equation (22.4.3.5) allows for the possibility that the conformation may change upon binding, since different charge distributions may be used for the complexed and uncomplexed forms of A, and similarly for B. However, other energetic terms, including those involved in any conformational change, have to be added to equation (22.4.3.5) to obtain net binding free energy changes. Nevertheless, changes in binding free energy due to charge modifications or changes in external factors such as pH and salt concentration may be estimated using equation (22.4.3.5) alone. For the latter, salt effects are usually only significant in highly charged molecules, for which the nonlinear form for the total electrostatic energy, equation (22.4.2.4), must be used. The salt dependence of binding of drugs and proteins to DNA has been studied using this approach (Misra, Hecht *et al.*, 1994; Misra, Sharp *et al.*, 1994; Sharp *et al.*, 1995), including the pH dependence of drug binding (Misra & Honig, 1995). Other applications include the binding of sulfate to the sulfate binding protein (Åqvist *et al.*, 1991) and antibody and antigen interactions (Lee *et al.*, 1992; Slagle *et al.*, 1994).

References

Antosiewicz, J., McCammon, J. A. & Gilson, M. K. (1994). *Prediction of pH-dependent properties of proteins. J. Mol. Biol.* **238**, 415–436.

Åqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). *Dipoles localized at helix termini of proteins stabilize charges. Proc. Natl Acad. Sci. USA*, **88**, 2026–2030.

Bacquet, R. & Rossky, P. (1984). *Ionic atmosphere of rodlike polyelectrolytes. A hypernetted chain study. J. Phys. Chem.* **88**, 2660.

Bashford, D. & Karplus, M. (1990). *pK$_a$'s of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry*, **29**, 10219–10225.

Beroza, P., Fredkin, D., Okamura, M. & Feher, G. (1991). *Protonation of interacting residues in a protein by a Monte-Carlo method. Proc. Natl Acad. Sci. USA*, **88**, 5804–5808.

Bharadwaj, R., Windemuth, A., Sridharan, S., Honig, B. & Nicholls, A. (1994). *The fast multipole boundary-element method for molecular electrostatics – an optimal approach for large systems. J. Comput. Chem.* **16**, 898–913.

Bruccoleri, R. E., Novotny, J., Sharp, K. A. & Davis, M. E. (1996). *Finite difference Poisson–Boltzmann electrostatic calculations: increased accuracy achieved by harmonic dielectric smoothing and charge anti-aliasing. J. Comput. Chem.* **18**, 268–276.

Davis, M. E. (1994). *The inducible multipole solvation model – a new model for solvation effects on solute electrostatics. J. Chem. Phys.* **100**, 5149–5159.

Davis, M. E. & McCammon, J. A. (1989). *Solving the finite difference linear Poisson Boltzmann equation: comparison of relaxational and conjugate gradient methods. J. Comput. Chem.* **10**, 386–395.

Gilson, M. (1993). *Multiple-site titration and molecular modeling: 2. Rapid methods for computing energies and forces for ionizable groups in proteins. Proteins Struct. Funct. Genet.* **15**, 266–282.

Gilson, M., Davis, M., Luty, B. & McCammon, J. (1993). *Computation of electrostatic forces on solvated molecules using the Poisson–Boltzmann equation. J. Phys. Chem.* **97**, 3591–3600.

Gilson, M. & Honig, B. (1986). *The dielectric constant of a folded protein. Biopolymers*, **25**, 2097–2119.

Gilson, M., McCammon, J. & Madura, J. (1995). *Molecular dynamics simulation with continuum solvent. J. Comput. Chem.* **16**, 1081–1095.

Gilson, M., Sharp, K. A. & Honig, B. (1988). *Calculating the electrostatic potential of molecules in solution: method and error assessment. J. Comput. Chem.* **9**, 327–335.

Holst, M., Kozack, R., Saied, F. & Subramaniam, S. (1994). *Protein electrostatics – rapid multigrid-based Newton algorithm for solution of the full nonlinear Poisson–Boltzmann equation. J. Biomol. Struct. Dyn.* **11**, 1437–1445.

Holst, M. & Saied, F. (1993). *Multigrid solution of the Poisson–Boltzmann equation. J. Comput. Chem.* **14**, 105–113.

Jayaram, B., Fine, R., Sharp, K. A. & Honig, B. (1989). *Free energy calculations of ion hydration: an analysis of the Born model in terms of microscopic simulations. J. Phys. Chem.* **93**, 4320–4327.

Jayaram, B., Sharp, K. A. & Honig, B. (1989). *The electrostatic potential of B-DNA. Biopolymers*, **28**, 975–993.

Jean-Charles, A., Nicholls, A., Sharp, K., Honig, B., Tempczyk, A., Hendrickson, T. & Still, C. (1990). *Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. J. Am. Chem. Soc.* **113**, 1454–1455.

Langsetmo, K., Fuchs, J. A., Woodward, C. & Sharp, K. A. (1991). *Linkage of thioredoxin stability to titration of ionizable groups with perturbed pK_a. Biochemistry*, **30**, 7609–7614.

Lee, F., Chu, Z. & Warshel, A. (1993). *Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the Polaris and Enzymix programs. J. Comput. Chem.* **14**, 161–185.

Lee, F. S., Chu, Z. T., Bolger, M. B. & Warshel, A. (1992). *Calculations of antibody–antigen interactions: microscopic and semi-microscopic evaluation of the free energies of binding of phosphorylcholine analogs to Mcpc603. Protein Eng.* **5**, 215–228.

Misra, V., Hecht, J., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on protein–DNA interactions: the lambda cI repressor and Eco R1 endonuclease. J. Mol. Biol.* **238**, 264–280.

Misra, V. & Honig, B. (1995). *On the magnitude of the electrostatic contribution to ligand–DNA interactions. Proc. Natl Acad. Sci. USA*, **92**, 4691–4695.

Misra, V., Sharp, K., Friedman, R. & Honig, B. (1994). *Salt effects on ligand–DNA binding: minor groove antibiotics. J. Mol. Biol.* **238**, 245–263.

Mohan, V., Davis, M. E., McCammon, J. A. & Pettitt, B. M. (1992). *Continuum model calculations of solvation free energies – accurate evaluation of electrostatic contributions. J. Phys. Chem.* **96**, 6428–6431.

Murthy, C. S., Bacquet, R. J. & Rossky, P. J. (1985). *Ionic distributions near polyelectrolytes. A comparison of theoretical approaches. J. Phys. Chem.* **89**, 701.

Nakamura, H., Sakamoto, T. & Wada, A. (1988). *A theoretical study of the dielectric constant of a protein. Protein Eng.* **2**, 177–183.

Nicholls, A. & Honig, B. (1991). *A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J. Comput. Chem.* **12**, 435–445.

Oberoi, H. & Allewell, N. (1993). *Multigrid solution of the nonlinear Poisson–Boltzmann equation and calculation of titration curves. Biophys. J.* **65**, 48–55.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1989). *Monte Carlo description of oligoelectrolyte properties of DNA oligomers. Proc. Natl Acad. Sci. USA*, **86**, 7766–7770.

Olmsted, M. C., Anderson, C. F. & Record, M. T. (1991). *Importance of oligoelectrolyte end effects for the thermodynamics of conformational transitions of nucleic acid oligomers. Biopolymers*, **31**, 1593–1604.

Pack, G., Garrett, G., Wong, L. & Lamm, G. (1993). *The effect of a variable dielectric coefficient and finite ion size on Poisson–Boltzmann calculations of DNA–electrolyte systems. Biophys. J.* **65**, 1363–1370.

Pack, G. R. & Klein, B. J. (1984). *The distribution of electrolyte ions around the B- and Z-conformers of DNA. Biopolymers*, **23**, 2801.

Pack, G. R., Wong, L. & Prasad, C. V. (1986). *Counterion distribution around DNA. Nucleic Acids Res.* **14**, 1479.

Rashin, A. A. (1990). *Hydration phenomena, classical electrostatics and the boundary element method. J. Phys. Chem.* **94**, 725–733.

Record, T., Olmsted, M. & Anderson, C. (1990). *Theoretical studies of the thermodynamics of ion interaction with DNA.* In *Theoretical Biochemistry and Molecular Biophysics.* New York: Adenine Press.

Reiner, E. S. & Radke, C. J. (1990). *Variational approach to the electrostatic free energy in charged colloidal suspensions. J. Chem. Soc. Faraday Trans.* **86**, 3901.

Schaeffer, M. & Frommel, C. (1990). *A precise analytical method for calculating the electrostatic energy of macromolecules in aqueous solution. J. Mol. Biol.* **216**, 1045–1066.

Sharp, K. & Honig, B. (1990). *Calculating total electrostatic energies with the non-linear Poisson–Boltzmann equation. J. Phys. Chem.* **94**, 7684–7692.

Sharp, K. A., Friedman, R., Misra, V., Hecht, J. & Honig, B. (1995). *Salt effects on polyelectrolyte–ligand binding: comparison of Poisson–Boltzmann and limiting law counterion binding models. Biopolymers*, **36**, 245–262.

Simonson, T. & Brooks, C. L. (1996). *Charge screening and the dielectric-constant of proteins: insights from molecular-dynamics. J. Am. Chem. Soc.* **118**, 8452–8458.

Simonson, T. & Brünger, A. (1994). *Solvation free energies estimated from a macroscopic continuum theory. J. Phys. Chem.* **98**, 4683–4694.

Simonson, T. & Perahia, D. (1995). *Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. Proc. Natl Acad. Sci. USA*, **92**, 1082–1086.

Sitkoff, D., Sharp, K. & Honig, B. (1994). *Accurate calculation of hydration free energies using macroscopic solvent models. J. Phys. Chem.* **98**, 1978–1988.

Slagle, S., Kozack, R. E. & Subramaniam, S. (1994). *Role of electrostatics in antibody–antigen association: anti-hen egg lysozyme/lysozyme complex (HyHEL-5/HEL). J. Biomol. Struct. Dyn.* **12**, 439–456.

Smith, P., Brunne, R., Mark, A. & van Gunsteren, W. (1993). *Dielectric properties of trypsin inhibitor and lysozyme calculated from molecular dynamics simulations. J. Phys. Chem.* **97**, 2009–2014.

Still, C., Tempczyk, A., Hawley, R. & Hendrickson, T. (1990). *Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc.* **112**, 6127–6129.

Takashima, S. & Schwan, H. P. (1965). *Dielectric constant measurements on dried proteins. J. Phys. Chem.* **69**, 4176.

Warshel, A. & Åqvist, J. (1991). *Electrostatic energy and macromolecular function. Annu. Rev. Biophys. Biophys. Chem.* **20**, 267–298.

Warshel, A. & Russell, S. (1984). *Calculations of electrostatic interactions in biological systems and in solutions. Q. Rev. Biophys.* **17**, 283.

Warwicker, J. (1994). *Improved continuum electrostatic modelling in proteins, with comparison to experiment. J. Mol. Biol.* **236**, 887–903.

Warwicker, J. & Watson, H. C. (1982). *Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. J. Mol. Biol.* **157**, 671–679.

Wendoloski, J. J. & Matthew, J. B. (1989). *Molecular dynamics effects on protein electrostatics. Proteins*, **5**, 313.

Yang, A., Gunner, M., Sampogna, R., Sharp, K. & Honig, B. (1993). *On the calculation of pK_a in proteins. Proteins Struct. Funct. Genet.* **15**, 252–265.

Yoon, L. & Lenhoff, A. (1992). *Computation of the electrostatic interaction energy between a protein and a charged suface. J. Phys. Chem.* **96**, 3130–3134.

Zauhar, R. & Morgan, R. J. (1985). *A new method for computing the macromolecular electric potential. J. Mol. Biol.* **186**, 815–820.

Zhou, H. X. (1994). *Macromolecular electrostatic energy within the nonlinear Poisson–Boltzmann equation. J. Phys. Chem.* **100**, 3152–3162.

International Tables for Crystallography (2012). Vol. F, Chapter 22.5, pp. 736–748.

Chapter 22.5. The relevance of the Cambridge Structural Database in protein crystallography

F. H. Allen, J. C. Cole and M. L. Verdonk

22.5.1. Introduction

At its inception in the late 1960s, the Cambridge Structural Database (CSD: Allen, Davies *et al.*, 1991; Kennard & Allen, 1993) was one of the first scientific databases for which numerical data were the primary objective of the compilation. Thus, the CSD provides not only a fully retrospective bibliography of the structure determination of organic and metallo-organic compounds, but also gives immediate access to the primary results of each diffraction experiment: the space group, cell dimensions and fractional coordinates that define each structure at atomic resolution. In the late 1960s, the world output of small-molecule structures was just a few hundred per year and it was possible to use existing printed compilations to ensure that the developing CSD was fully retrospective. Despite this comprehensive nature, it has taken time for the CSD to have significant scientific impact as a research tool in its own right, and to be recognized as a source of structural knowledge that is applicable across a broad spectrum of structural chemistry.

There are two reasons for this rather gradual uptake. First, it took time to devise and implement software for the validation and organization of the data. Secondly, and most importantly, it was necessary to develop software for database searching, particularly for locating chemical substructures, and for data analysis and visualization. It was not until the late 1970s that the first comprehensive software systems became available and began to be widely distributed to scientists in academia and industry. Nevertheless, a number of highly influential database analyses were performed prior to 1980, and the proper numerical analysis and statistical treatment of bulk geometrical data began to receive attention (see *e.g.* Murray-Rust & Bland, 1978; Murray-Rust & Motherwell, 1978; Taylor, 1986). This software and its successors at last allowed the types of geometrical surveys, analyses and tabulations carried out manually by early practitioners such as Pauling (1939), Sutton (1956, 1959) and Pimental & McClellan (1960) to be executed automatically in a few minutes of increasingly powerful CPU time.

The early development of applications software simultaneously with methods for the acquisition and validation of new structural data was crucial for the CSD. Developments in structure-determination theory, allied to technological improvements in data collection and the ever increasing speed and capacity of modern computers, led to such a rapid expansion that the archive of May 1999 now contains more than 200 000 crystal structures, a total that doubles approximately every seven years. The literature is now so vast, so chemically diverse and so widely spread that it is virtually impossible for individual scientists to maintain current awareness without recourse to database facilities. It is now impossible to carry out viable systematic analyses without recourse to database technology. This chapter focuses primarily on the structural knowledge that is provided by such analyses, and that is relevant to the determination, refinement, validation and systematic study of macromolecular structures. However, the validity of these results depends crucially on two factors: the *completeness* of the archive and the *accuracy* with which the data are recorded. Hence, it is appropriate to preface the chapter with some comparative comment on these fundamental issues as they apply to the small-molecule and macromolecular structure archives.

22.5.2. The CSD and the PDB: data acquisition and data quality

22.5.2.1. Statistical inferences

With a current total of 200 000 structures and a doubling period of seven years (Fig. 22.5.2.1*a*), we may expect at least half a million small-molecule crystal structures to be in the CSD by the year 2010. The Protein Data Bank (PDB) (Abola *et al.*, 1997; Berman *et al.*, 2000), which began operations in the mid-1970s, and the Nucleic Acid Database (NDB) (Berman *et al.*, 1992) are the international repositories for macromolecular structure information. Input to the PDB was initially slow but is now showing a rapid growth rate reminiscent of the CSD of the 1970s (Fig. 22.5.2.1*b*). The PDB archive has a current total of *ca* 8500 structures (mid-1999) and a doubling period of close to two years. As with the CSD, this early *high rate* of growth will almost certainly decrease, thus increasing the doubling period. Nevertheless, by the year 2010, we might expect the PDB to contain more than 100 000 structures.

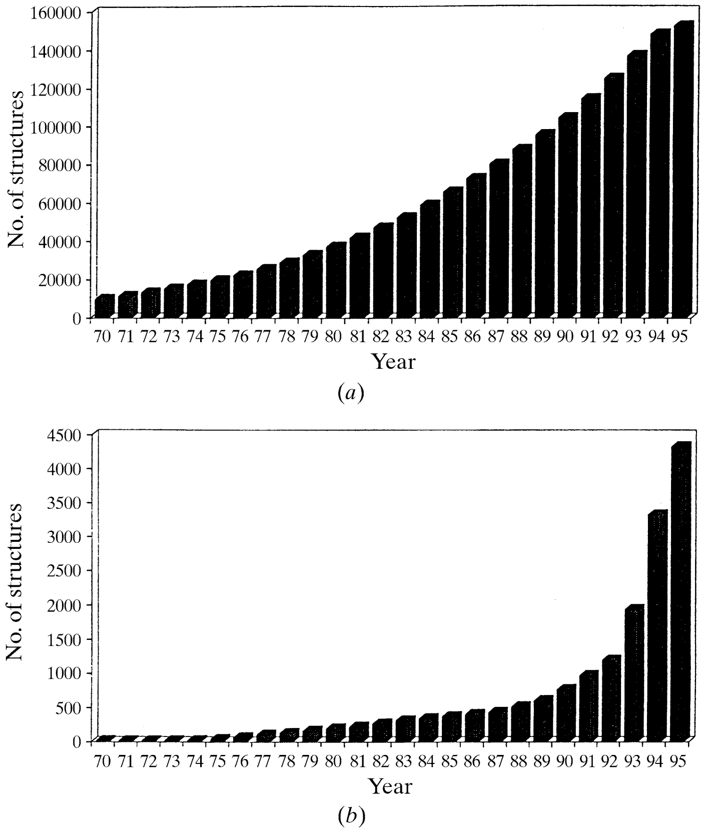


Figure 22.5.2.1
(*a*) Growth rate of the CSD and (*b*) growth rate of the PDB, in terms of the numbers of structures published per annum for the period 1970–1995.

Copyright © 2012 International Union of Crystallography

22.5.2.2. Data acquisition and completeness

Given the size and diversity of the CSD, it is amazing that searches for some common chemical substructures often yield far fewer hits than might have been expected. Sometimes, the absence of just a few key CSD entries would have negated a successful systematic analysis: some points in a graph would have been missing and a correlation would not have been detected. Similarly, completeness of the PDB is vital for the future of 'data mining' or 'knowledge engineering' in the macromolecular arena.

Data acquisition by the PDB has always had one valuable advantage in comparison with the CSD. The volume of numerical data generated by a protein structure determination is far too large for primary publication or hard-copy deposition. Thus, the PDB has always acquired data through direct deposition in electronic form, and authors have usually been involved in the validation of their entries. Further, it is a mandatory requirement of the vast majority of journals, and a clear recommendation of appropriate professional organizations, that prior deposition with the PDB is an essential precursor to primary publication. This key involvement of the PDB in the publication process acts as a vital guarantee of the completeness of the archive. The prior-deposition rule must be rigidly adhered to for the long-term benefit of science.

22.5.2.3. Standard formats: CIF and mmCIF

The CSD, on the other hand, reflects the published literature, and much of its data content has been re-keyboarded from hard-copy material. The Cambridge Crystallographic Data Centre (CCDC) is now beginning to receive significant amounts of electronic input, a development that owes much to the rapid international acceptance of an agreed standard electronic interchange format, the crystallographic information file or CIF (Hall *et al.*, 1991), and the rapid incorporation of CIF generators within most major structure solution and refinement packages. The CIF offers many advantages, some of which are only just being addressed within the CSD: (*a*) a clear definition of input data items and their representation; (*b*) a significant reduction in time spent correcting simple typographical errors; and (*c*) the possibility of enhancing the overall database content through the electronic availability of *all* information from the analysis, *i.e.* more than could reasonably be re-typed from hard-copy material. For the PDB, the recent adoption of the macromolecular CIF (mmCIF) as the agreed international standard offers similar advantages. This development, together with advances in communications technology, now make it possible to automate the deposition process more effectively, but the advantages of mmCIF can only be fully realized once it also becomes a standard output format of all of the relevant software packages.

22.5.2.4. Structure validation

The value of research results derived from the CSD and the PDB depends crucially on the accuracy of the underlying data [see *e.g.* Hooft *et al.* (1996) with respect to protein data]. As with the early CSD, much current research involves use of data from the developing PDB to establish rules and protocols for the validation of new protein structures (see *e.g.* Laskowski *et al.*, 1993). This activity, in turn, means that earlier entries in the archive may have to be reassessed periodically to bring their representations into line with best current practice. This sequence of events was commonplace in the CSD of the 1970s and, even now, new structure types entering the CSD can still provoke a reassessment of subclasses of earlier entries.

Secondly, it is important that errors and warnings raised by validation software have clear meanings and that validation results are clearly encoded within each entry. The end user can then make informed choices about which entries to include (or not) in any given application. Recent moves to apply a range of agreed and unambiguous primary checks to new data, and to require resolution of any problems prior to the issue of a publication ID code, represent an important development.

22.5.3. Structural knowledge from the CSD

22.5.3.1. The CSD software system

Structural knowledge from the CSD is reflected principally in the geometries of individual molecules, extended crystal structures and, most importantly, through systematic studies of the geometrical characteristics of large subsets of related substructural units. Software facilities for search, retrieval, analysis and visualization of CSD information as of 2001 are fully described in Chapter 24.3 of the first edition of this volume. The system allows for the calculation of a very wide range of geometrical parameters, both intramolecular and intermolecular. Most importantly, chemical substructural search fragments may be specified using normal covalent bonding definitions (single, double, triple *etc.*), limiting non-covalent contact distances and other geometrical constraints. For each instance of a search fragment located in the CSD, the system will compute a user-defined set of geometrical descriptors. The full matrix, $G(N, p)$, of the p geometrical parameters for each of the N fragments located in the CSD can then be analysed using numerical, statistical and visualization techniques to display individual parameter distributions, to compute medians, means and standard deviations, and to examine the geometrical data for correlations or discrete clusters of observations that may exist in the p-dimensional parameter space.

22.5.3.2. CSD structures and substructures of relevance to protein studies

Table 22.5.3.1 presents statistics for the 3137 structures of amino acids and peptides that are available in the CSD of April 1998 (containing 181 309 entries). Although this represents less than 2% of CSD information, some may consider that these are the only entries of real interest in molecular biology. In certain cases, *e.g.* for the derivation of very precise molecular dimensions and for some conformational work, this may be true. However, the real issue concerns the *transferability* of CSD-derived information to the protein environment. It is the biological relevance of a chemical *substructure* (inter- or intramolecular) that is important, and this consideration immediately brings much larger subsets of CSD entries into play. Information such as van der Waals radii can be derived from the CSD as a whole, while more specific information concerning, for example, biologically important metal coordination geometries can be derived from appreciable subsets of the total database, as shown in the statistics of Table 22.5.3.2.

22.5.3.3. Geometrical parameters of relevance to protein studies

Precise geometrical knowledge from atomic resolution studies of small molecules is important in the macromolecular domain since it provides: (*a*) geometrical restraints and standards to be applied during protein structure determination, refinement and validation; (*b*) model geometries for liganded small molecules and information about their preferred modes of interaction with

Table 22.5.3.1
Summary of amino-acid and peptide structures available in the CSD (April 1998, 181 309 entries)

(*a*) Overall statistics

Structures	No. of entries
α-Amino acids (any organic) †	3137
Peptides (standard or modified standard α-amino acids) ‡	1430

(*b*) Peptide statistics

	No. of CSD entries	
No. of residues	Acyclic	Cyclic
2	543	123
3	249	45
4	76	50
5	62	44
6	20	73
7	14	15
8	19	32
10	16	19
11	4	10
12	2	11
13	—	—
14	1	—
15	3	2
16	3	—

† Any organic structure containing the α-amino acid functionality. ‡ The standard amino acids (those normally found in proteins) may be modified by substitution in these peptides.

Table 22.5.3.2
CSD entry statistics for selected metal-containing structures

CSD entries ($R < 0.10$) containing M and (N or O). No additional transition metals were allowed to occur in the Na, K, Mg and Ca structures cited.

Metal	No. of CSD entries
Na	1189
K	987
Mg	510
Ca	469
Zn	1996

the host protein; (*c*) details of metal coordination spheres and geometries that are likely to be observed in metalloproteins; and (*d*) information from which force field and other parameters may be derived. Thus, the types of study discussed in this chapter are concerned with retrieving systematic knowledge concerning:

(1) molecular dimensions: bond lengths and valence angles;
(2) conformational features: torsion angles that describe acyclic and cyclic systems;
(3) metal coordination-sphere geometries: coordination numbers, metal–ligand distances and inter-ligand valence angles;
(4) general non-bonded contact distances: van der Waals radii;
(5) hydrogen-bond geometries: distances, angles, directional properties;
(6) other non-bonded interactions: identification and geometrical description;
(7) formation of preferred atomic arrangements or motifs involving non-covalent interactions.

In this short overview, which deals with such a broad range of structural information, our literature coverage is, of necessity, highly selective. In each area, we have tried to cite the more recent papers, from which leading references to earlier studies can be located. We also draw attention to a number of recent monographs in which a variety of CSD analyses are comprehensively cited and discussed: *Structure Correlation* (Bürgi & Dunitz, 1994), *Crystal Structure Analysis for Chemists and Biologists* (Glusker *et al.*, 1994), *Hydrogen Bonding in Biological Structures* (Jeffrey & Saenger, 1991) and *Crystal Engineering: the Design of Organic Solids* (Desiraju, 1989). Finally, we note the CCDC's own database of published research applications of the CSD. The DBUSE database currently contains literature references and short descriptive abstracts for nearly 700 papers. It forms part of each biannual CSD release and is fully searchable using the *Quest3D* program.

22.5.4. Intramolecular geometry

22.5.4.1. Mean molecular dimensions

The work of Pauling (1939) represented the first systematic attempt to derive mean values for bond lengths and valence angles from the limited structural data available at that time. This work resulted in the definition of covalent bonding radii for the common elements and had a seminal influence on the development of chemistry over the past half century. Further tabulations appeared sporadically until the publication in 1956 and 1959 of the major compilation *Tables of Interatomic Distances and Configuration in Molecules and Ions,* edited by Sutton (1956, 1959), by The Chemical Society of London. Kennard (1962) extended the available data for bonds between carbon and other elements.

In the mid-1980s, the CCDC and its collaborators compiled updated tables of mean bond lengths for both organic (Allen *et al.*, 1987) and organometallic and metal coordination compounds (Orpen *et al.*, 1989). Both compilations were based on the CSD of September 1985 containing 49 854 entries. Of these, 10 324 organic structures and 9802 organometallics or metal complexes satisfied a variety of secondary selection criteria, and were used in the analysis. For each bond length, both compilations present the mean, its estimated standard deviation and the sample standard deviation, together with the median value of the distribution and its upper and lower quartile values. The organic section describes 682 discrete chemical bond types involving 65 element pairs. Of these, 511 (75%) involve carbon, and 428 (63%) involving only carbon, nitrogen and oxygen are relevant to protein studies. The organometallic and metal complex compilation presents similar statistics for 325 different bond types involving *d*- and *f*-block metals. It is planned to automate and systematize the production of such tabulations, so that they can be dynamically updated in computerized form, as part of CCDC's ongoing development of knowledge-based structural libraries.

More recently, Engh & Huber (1991) have generated sets of mean bond lengths and valence angles from peptidic structures retrieved from the 80 000 entries then available in the CSD. Their compilations are based on 31 atom types which are most appropriate to the protein environment and are well represented in CSD structures. These authors note that such knowledge, together with torsional and other information, is vital to the determination, refinement and validation of protein structures. Prior to their detailed CSD analysis, some of the parameters used for these purposes had been determined with a lower accuracy than was required by the diffraction data. For this reason, and particularly for use with higher-resolution protein data, they recommend that the most accurate parameters possible should always be used.

Systematic use of CSD data generates mean values together with standard deviations for both the sample and the mean. The sample standard deviations provide information about the spread of each parameter distribution, *i.e.* information about the variability of each parameter which can be parameterized as force constants. Comparative refinements of selected proteins showed that the new CSD-based parameters yielded significant improvements in R factors and in geometry statistics. Finally, Engh & Huber (1991) remark that their results should be updated regularly as the quantity and quality of data in the CSD increase with time. Apart from producing more precise estimates of mean values, incorporation of more protein-relevant atom types into the schema should then be possible.

22.5.4.2. Conformational information

Torsion angles are the natural measures of conformational relationships within molecules. If we specify a chemical substructure involving a central bond of interest, then the CSD system will display the distribution of torsion angles about that bond, computed from the tens, hundreds, or even thousands of instances located in the database. Examination of these *univariate distributions* will reveal any conformational preferences that may exist in small-molecule crystal structures. This approach is illustrated by the histogram of Fig. 22.5.4.1, which shows the torsional distribution about S—S bridge bonds in C(sp^3)—S—S—C(sp^3) substructures located in the CSD. Clearly, there is a preference for a perpendicular conformation in the CS—SC unit. This corresponds well with values observed for cysteine bridges in protein structures, and with theoretical calculations on small model compounds.

The interrelationship between two torsion angles can be visualized by plotting them against each other on a conventional 2D scattergram. In the small-molecule area, the distribution of data points in these scattergrams can reveal conformational interconversion pathways (Rappoport *et al.*, 1990) or show areas of high data density corresponding to conformational preferences (Schweizer & Dunitz, 1982). The best known *bivariate distribution* is the Ramachandran plot of peptidic φ–ψ angles, which is universally used to assess the quality of protein structures and to identify structural features. Ashida *et al.* (1987) performed an extensive analysis of peptide conformations available in the CSD and present torsional histograms, a Ramachandran plot, and a variety of other visual and descriptive statistics that summarize this data set.

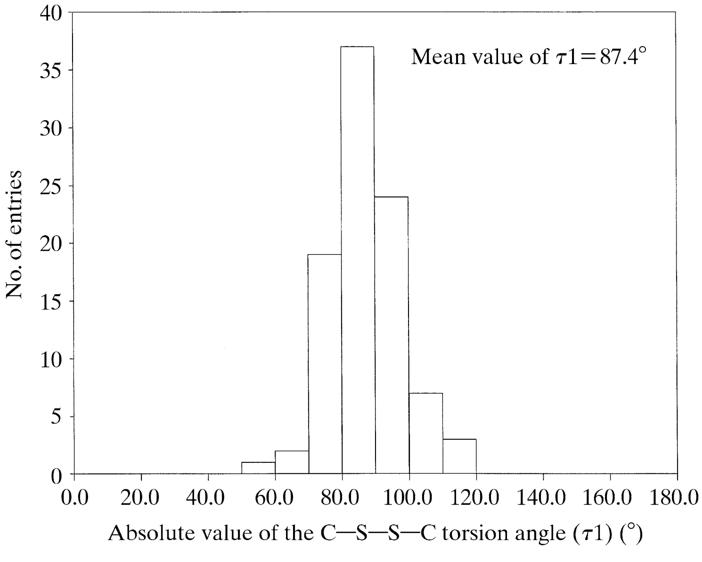


Figure 22.5.4.1
Distribution of torsion angles in C(sp^3)—S—S—C(sp^3) substructures located in the CSD.

It is often necessary to use three or more torsion angles to define the conformation of, *e.g.*, a side chain or flexible ring. Here, *multivariate* statistical techniques (Chatfield & Collins, 1980; Taylor, 1986) have proved valuable for extracting information from the matrix $T(N, k)$ that contains the k torsion angles computed for each of the N examples of the substructure in the CSD. Two methods, both available within the CSD system software described in Chapter 24.3 of the first edition of this volume, are commonly used to visualize the k-dimensional data set and to locate natural sub-groupings of data points within it.

Principal component analysis (PCA) (Murray-Rust & Motherwell, 1978; Allen, Doyle & Auf der Heyde, 1991; Allen, Howard & Pitchford, 1996) is a dimension-reduction technique which analyses the variance in $T(N, k)$ in terms of a new set of uncorrelated, orthogonal variables: the principal components, or PCs. The PCs are generated in decreasing order of the percentage of the variance that is explained by each of them. The hope is that the number of PCs, p, that explains most of the variance in the data set is such that $p \ll k$, so that a few pairwise scatter plots with respect to the new PC axes will provide useful visualizations of the complete data set. For cyclic fragments, PCA results are closely related to those obtained using the ring-puckering methodology of Cremer & Pople (1975). Cluster analysis (CA) (Everitt, 1980; Allen, Doyle & Taylor, 1991) is a purely numerical method that attempts to locate discrete groupings of data points within a multivariate data set. CA uses 'distances' or 'dissimilarities' between pairs of points in a k-dimensional space as its working basis, and a very large number of clustering algorithms exist. The mathematical basis of both of these techniques, the modifications that are needed to account for topological symmetry in the search fragment and examples of their application have been reviewed by Taylor & Allen (1994).

Preliminary work using the concepts of machine learning (Carbonell, 1989) for knowledge discovery and classification have also been carried out using the CSD (see *e.g.* Allen *et al.*, 1990; Fortier *et al.*, 1993). In particular, conceptual clustering methods have been applied to a number of substructures (Conklin *et al.*, 1996) and the results compared with those obtained by the statistical and numerical methods described above. Similar techniques are also being used for the classification of protein structures (see *e.g.* Blundell *et al.*, 1987).

22.5.4.3. Crystallographic conformations and energies

Crystallographic conformations obviously represent energetically accessible forms. However, for use in molecular-modelling applications, the key question must be asked: Are the condensed-phase crystallographic observations a good guide to conformational preferences in other phases? The indications are that the answer is 'yes' from the types of studies exemplified or cited in the previous section: there appears to be a clear *qualitative* relationship between crystallographic conformer distributions and the low-energy features of the appropriate potential energy hypersurface, although the estimation of absolute energies from the relative populations of these distributions is not appropriate (Bürgi & Dunitz, 1988).

Allen, Harris & Taylor (1996) addressed this question in a systematic manner for a series of 12 one-dimensional (univariate) conformational problems. All of the chosen substructures [simple derivatives of ethane, involving a single torsion angle (τ) about the central C—C bond] were expected to show one symmetric (*anti*, $\tau \simeq 180^\circ$) energy minimum and two symmetry-related asymmetric (*gauche*, $\tau \simeq \pm 60^\circ$) minima. For each substructure, the crystallographic torsional distribution was determined from the CSD and compared with the 1D potential-energy profile, computed using *ab initio* molecular-orbital methods and the 6–31G* basis set. Close agreement was observed between the experimental condensed phase results and the computed *in vacuo* data. Taken over all 12 substructures, the *ab initio* optimized values of the asymmetric (*gauche*) torsion angle vary from <55° to >80°, and a scatter plot of these optimized values *versus* the mean crystallographic values for *gauche* conformers is linear, with a correlation coefficient of 0.831. Two other results of the study were that: (*a*) torsion angles with higher strain energies (>4.5 kJ mol^{-1}) are rarely observed in crystal structures (<5%); and (*b*) taken over many structures, conformational distortions due to crystal packing appear to be the exception rather than the rule.

22.5.4.4. *Conformational libraries*

In essence, the CSD can be regarded as a huge library of individual molecular conformations. However, to be of general value, it is necessary to distil, store and present this knowledge in an ordered manner, in the form of torsional distributions for specific atomic tetrads $A—B—C—D$. Protein-specific libraries of this type derived from high-resolution PDB structures are commonly used as aids to protein structure determination, refinement and validation (Bower *et al.*, 1997; Dunbrack & Karplus, 1993). The information can either be stored in external databases, or hardwired into the program in the form of rules. However, CSD usage has tended to concentrate on analyses of individual substructures, as noted above, both for their intrinsic interest and to develop novel methods of data analysis. Recently, Klebe & Mietzner (1994) have described the generation of a small library containing 216 torsional distributions derived from the CSD, together with 80 determined from protein–ligand complexes in the PDB. The library was used in a knowledge-based approach for predicting multiple conformer models for putative ligands in the computational modelling of protein–ligand docking. Conformer prediction is accomplished by the computer program *MIMUMBA*. As part of its programme for the development of knowledge-based libraries from the CSD, the CCDC has now embarked on the generation of a more comprehensive torsional library. Here, information is being hierarchically ordered according to the level of specificity of the chemical substructures for which torsional distributions are available in the library.

22.5.4.5. *Metal coordination geometry*

Some 54% of the information content of the CSD relates to organometallics and metal complexes. This reflects the crucial role of single-crystal diffraction analyses in the renaissance of inorganic chemistry since the 1950s, and the fundamental importance of the technique in characterizing the many novel molecules synthesized over the past 40 years. Since ligands containing nitrogen, oxygen and sulfur are ubiquitous, the CSD contains much information that is relevant to the binding of metal ions by proteins [*e.g.* zinc (Miller *et al.*, 1985), calcium (Strynadka & James, 1989) *etc.*]. Some statistics for the occurrence of some common metals having N and/or O ligands are presented in Table 22.5.3.2.

One of the earliest studies (Einspahr & Bugg, 1981) concerned the geometry of Ca–carboxylate binding, with special reference to biological systems. Since that time, a variety of other studies of biologically relevant metal coordination modes have appeared from the laboratories of Glusker, Dunitz and others (see *e.g.* Glusker, 1980; Chakrabarti & Dunitz, 1982; Carrell *et al.*, 1988, 1993; Chakrabarti, 1990*a*,*b*). These studies show, *inter alia*, that α-hydroxycarboxylates and imidazoles such as histidine tend to bind metal ions in their planes, but that alkali metal cations tend to bind carboxylate groups indiscriminately both in-plane and out-of-plane. Chapter 17 of Glusker *et al.* (1994) is a significant source of additional information and leading references to work in this area over the past two decades.

22.5.5. Intermolecular data

Non-bonded interaction geometries observed in small-molecule crystal structures are of great value in the determination and validation of protein structures, in furthering our understanding of protein folding, and in investigating the recognition processes involved in protein–ligand interactions. The CSD continues to provide vital information on all of these topics.

22.5.5.1. *van der Waals radii*

The hard-sphere atomic model is central to chemistry and molecular biology and, to an approximation, atomic van der Waals radii can be regarded as transferable from one structure to another. They are heavily used in assessing the general correctness of all crystal-structure models from metals and alloys to proteins. Pauling (1939) was the first to provide a usable tabulation for a wide range of elements, but the values of Bondi (1964) remain the most highly cited compilation in the modern literature. His values, assembled from a variety of sources including crystal-structure information, were selected for the calculation of molecular volumes and, in his original paper, Bondi (1964) issues a caution about their general validity for the calculation of limiting contact distances in crystals. In view of the huge amount of non-bonded contact information available in the CSD, Rowland & Taylor (1996) recently tested Bondi's statement as it might apply to the common nonmetallic elements, *i.e.* H, C, N, O, F, P, S, Cl, Br and I. They found remarkable agreement (within 0.02 Å) between the crystal-structure data and the Bondi values for S and the halogens, and agreement within 0.05 Å for C, N and O (new values all larger). The only significant discrepancy was for H, where averaged neutron-normalized small-molecule data yield a van der Waals radius of 1.1 Å, 0.1 Å shorter than the Bondi (1964) value. In the specific area of amino-acid structure, Gould *et al.* (1985) have studied the crystal environments and geometries of leucines, isoleucines, valines and phenylalanines. Their work provides estimates of minimum non-bonded contact distances and indicates the preferred van der Waals interactions of these primary building blocks.

22.5.5.2. *Hydrogen-bond geometry and directionality*

The hydrogen bond is the strongest and most frequently studied of the non-covalent interactions that are observed in crystal structures. As with intramolecular geometries, the first surveys of non-bonded interaction geometries all concerned hydrogen bonds, and were reported long before the CSD existed (Pauling, 1939; Donohue, 1952; Robertson, 1953; Pimentel & McClellan, 1960). The review by Donohue (1952) already contained a plot of N···O distances *versus* C—N···O angles in crystal structures (the C—N groups are terminal charged amino groups), while the review by Pimentel & McClellan (1960)

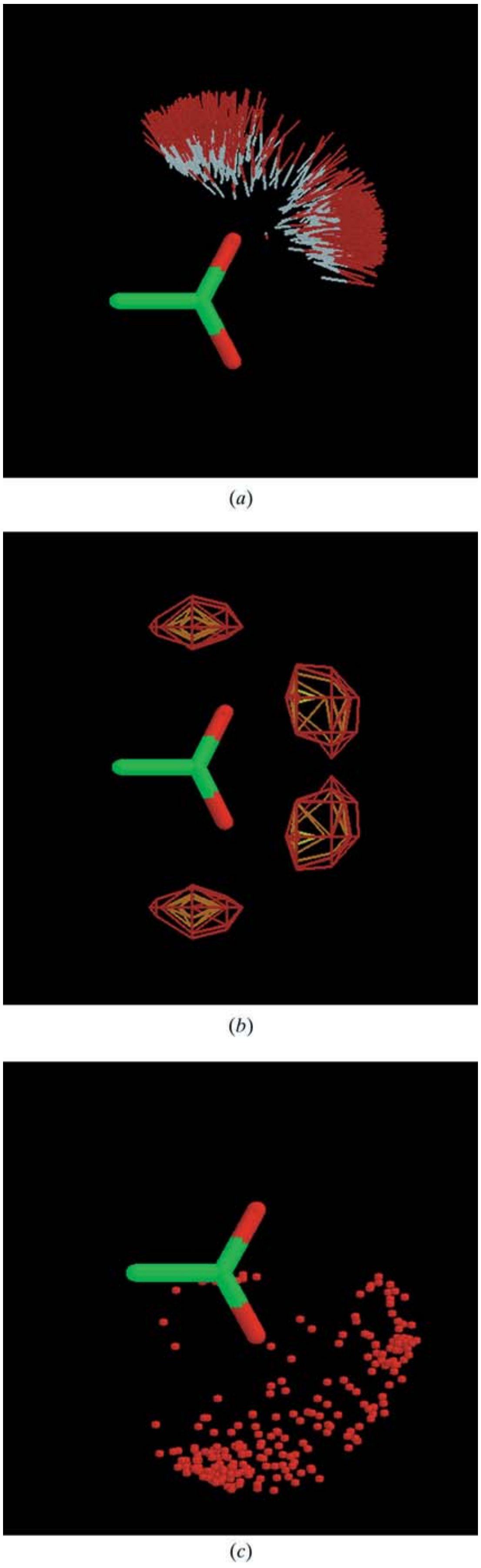


Figure 22.5.5.1
The IsoStar knowledge-based library of intermolecular interactions: interaction of O—H donors (contact groups) with one of the >C═O acceptors of a carboxylate group (the central group). (*a*) Direct scatter plot derived from CSD data, (*b*) contoured scatter plot derived from CSD data and (*c*) direct scatter plot derived from PDB data.

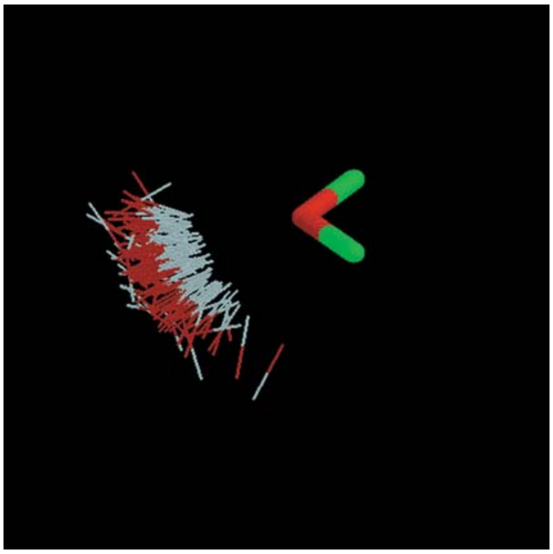


Figure 22.5.5.2
Distribution of O—H donors around ether oxygen acceptors (CSD data from the IsoStar library, see text).

contained histograms of hydrogen-bond distances. Up to the mid-1970s, numerous other studies appeared, *e.g.* Balasubramanian *et al.* (1970), Kroon & Kanters (1974) and Kroon *et al.* (1975), in which all of the statistical analyses were performed manually.

With the advent of the CSD and its developing software system, these kinds of studies became much more accessible and easier to perform, although the non-bonded search facility was only generalized and fully integrated within *Quest3D* in 1992. Thus, Taylor and colleagues reported studies on N—H···O═C hydrogen bonds (Taylor & Kennard, 1983; Taylor *et al.*, 1983, 1984*a*,*b*), Jeffrey and colleagues reported detailed studies on the O—H···O hydrogen bond (Ceccarelli *et al.*, 1981), hydrogen bonds in amino acids (Jeffrey & Maluszynska, 1982; Jeffrey & Mitra, 1984), and hydrogen bonding in nucleosides and nucleotides, barbiturates, purines and pyrimidines (Jeffrey & Maluszynska, 1986), while Murray-Rust & Glusker (1984) studied the directionalities of O—H···O hydrogen bonds to ethers and carbonyls. These studies indicated that hydrogen bonds are often very directional. For example, the distribution of the O—H···O hydrogen-bond angle, after correction for a geometrical factor, peaks at 180° (*i.e.* there is a clear preference for linear hydrogen bonds) and, in carbonyls and carboxylate groups, hydrogen bonds tend to form along the lone-pair directions of the O-atom acceptors (Fig. 22.5.5.1). For ethers, however, lone-pair directionality is not observed, as is illustrated in Fig. 22.5.5.2.

Software availability has facilitated CSD studies of a wide range of individual hydrogen-bonded systems in the recent literature, including studies of resonance-assisted hydrogen bonds (Bertolasi *et al.*, 1996) and resonance-induced hydrogen bonding to sulfur (Allen, Bird *et al.*, 1997*a*). These statistical studies are often combined with molecular-orbital calculations of interaction energies. Some of these studies are cited in this chapter, but the monograph of Jeffrey & Saenger (1991) and the CCDC's DBUSE database are valuable reference sources.

22.5.5.3. *C—H···X hydrogen bonds*

An important and often underestimated interaction in biological systems is the C—H···*X* hydrogen bond. These bonds have been extensively studied in small-molecule crystal structures, especially in relation to the ongoing discussion as to whether or not they should be called hydrogen bonds. Although Donohue (1968) concluded that the question 'The C—H···O hydrogen bond: what is it?' had only one answer: 'It isn't', a survey of 113 neutron-diffraction structures showed clear statistical evidence for an attractive interaction between C—H groups and oxygen and nitrogen acceptors (Taylor & Kennard, 1982). Later, more evidence for this hypothesis was found, and it was even shown that some C—H···O interactions are directional (Berkovitch-Yellin & Leiserowitz, 1984; Desiraju, 1991; Steiner & Saenger, 1992; Desiraju *et al.*, 1993; Steiner *et al.*, 1996). A continuing area of interest has been to establish the relative donor abilities of C—H in different chemical environments, since spectroscopic data had indicated that donor ability decreased in the order $C(sp)$—H > $C(sp^2)$—H > $C(sp^3)$—H. This general hydrogen-acidity requirement was noted by Taylor & Kennard (1982), and systematically addressed using CSD information by Desiraju & Murty (1987), and by Pedireddi & Desiraju (1992), who derived a novel scale of carbon acidity based on C···O separations in a wide variety of systems containing C—H···O hydrogen bonds. A recent paper (Derewenda *et al.*, 1995) highlights the importance of C—H···O═C bonds in stabilizing protein secondary structure.

22.5.5.4. O—H···π and N—H···π hydrogen bonds

Spectroscopic evidence for the existence of N,O—H···π hydrogen bonding to acetylenic, olefinic and aromatic acceptors is well documented (Joris *et al.*, 1968). To our knowledge, the first survey of these interactions in the CSD was carried out by Levitt & Perutz (1988), prompted by observations made in protein structures. A more recent CSD survey of this type of bonding (Viswamitra *et al.*, 1993) has shown that intermolecular examples are clearly observed and that these bonds, although very weak, can be both structurally and energetically significant. Recently, Steiner *et al.* (1995) have presented novel crystal structures, database evidence and quantum-chemical calculations on C≡C—H···π(C≡C) and π(phenyl) bonding. They cite H···C≡C (midpoint) distances as short as 2.51 Å and observe hydrogen-bond cooperativity in extended systems with hydrogen-bond energies in the range 4.2–9.2 kJ mol^{-1}. Finally, we note that electron-rich transition metals can act as proton acceptors in hydrogen-bond interactions with O—H, N—H and C—H donors. Brammer *et al.* (1995) have reviewed progress in this developing area.

22.5.5.5. Other non-covalent interactions

The hydrogen bond, $X(\delta-)$—$H(\delta+)\cdots Y(\delta-)$—$Z(\delta+)$, can be viewed as an (almost) linear dipole–dipole interaction, whose ubiquity in nature is due to the presence of many donor–hydrogen dipoles. In a recent review of supramolecular synthons and their application in crystal engineering, Desiraju (1995) illustrates the structural importance of a wide range of attractive non-bonded interactions that do not involve hydrogen mediacy, and notes the long-term value of the CSD in identifying and characterizing these interactions. The area of weak intermolecular interactions is now a burgeoning one in which the combination of CSD analysis and high-level *ab initio* molecular-orbital calculations is proving important in establishing both preferred geometries and estimates of interaction energies. In this context, the intermolecular perturbation theory (IMPT) of Hayes & Stone (1984), a methodology which is free of basis-set superposition errors, is proving particularly useful.

Some of the earliest CSD studies concerned the geometry and directionality of approach of N and O nucleophiles to carbonyl centres, leading to the mapping of (dynamic) reaction pathways through systematic analysis of many examples of related (static) crystal structures (see Bürgi & Dunitz, 1983, 1994). This work was also extended to a study of the directional preferences of non-bonded atomic contacts at sulfur atoms, initially using S in amino acids but later including other examples of divalent sulfur (Rosenfield *et al.*, 1977). It was shown that C—S—C groups tend to bind positively charged electrophiles in directions that are approximately perpendicular to the C—S—C plane, while negatively charged nucleophiles prefer to bind to S along an extension of one of the C—S bonds.

The strong tendency for halogens X = Cl, Br and I to form short contacts to other halogens, and especially to electronegative O and N atoms (Nyburg & Faerman, 1985) is well known (Price *et al.*, 1994). Recent combined CSD/IMPT studies of C—X···O═C (Lommerse *et al.*, 1996) and C—X···O(nitro) (Allen, Lommerse *et al.*, 1997) systems showed a marked preference for the X···O interaction to form along the extension of the C—X bond, with interaction energies in the range −7 to −10 kJ mol^{-1}. These interactions have been used (Desiraju, 1995) to engineer a variety of novel small-molecule crystal structures, and the few X···O interactions observed in protein structures generally conform to the geometrical preferences observed in small-molecule studies.

Interactions involving other functional groups are also of importance, and Taylor *et al.* (1990) used CSD information to construct composite crystal-field environments for carbonyl and nitro groups in their search for isosteric replacements in modelling protein–ligand interactions. Their work showed that many of the short intermolecular contacts made by carbonyl groups are to other carbonyl groups in the extended crystal structure. More recently, Maccallum *et al.* (1995*a*,*b*) have demonstrated the importance of Coulombic interactions between the C and O atoms of proximal CONH groups in proteins as an important factor in stabilizing α-helices, β-sheets and the right-hand twist often observed in β-strands. Their calculations indicate an attractive carbonyl–carbonyl interaction energy of about −8 kJ mol^{-1} in specific cases, and they remark that these interactions are *ca* 80% as strong as the CO···HN hydrogen bonds within their computational model. Allen, Baalham *et al.* (1998) have used combined CSD/IMPT analysis in a more detailed study of carbonyl–carbonyl interactions and have shown that (*a*) the interaction is commonly observed in small-molecule structures; (*b*) that the preferred interaction geometry is a dimer motif involving two antiparallel C···O interactions, although numerous examples of a perpendicular motif (one C···O interaction) were also observed; and that (*c*) the total interaction energies for the antiparallel and perpendicular motifs are about −20 and −8 kJ mol^{-1}, respectively, the latter value being comparable to that computed by Maccallum *et al.* (1995*a*,*b*). In studies with protein structures, it has also been noted that carbonyl–carbonyl interactions stabilize the partially allowed Ramachandran conformations of asparagine and aspartic acid (Deane *et al.*, 1999)

22.5.5.6. Intermolecular motif formation in small-molecule crystal structures

Desiraju (1995) has stressed that the design process in crystal engineering depends crucially on the high probabilities of formation of certain well known intermolecular motifs, *e.g.* the hydrogen-bonded dimer frequently formed by pairs of carboxylate groups. By analogy with molecular synthesis, he describes these general non-covalent motifs (which often contain strong hydrogen bonds) as supramolecular *synthons*, and points to their importance in supramolecular chemistry as a whole (see *e.g.* Lehn, 1988; Whitesides *et al.*, 1995). Since protein–protein and protein–ligand interactions are also supramolecular phenomena, it follows that information about common interaction motifs is also of importance in structural biology. A computer program is now being written at the CCDC to establish the topologies, chemical constitutions and probabilities of formation of intermolecular motifs directly from the CSD. Initial results (Allen, Raithby *et al.*, 1998; Allen *et al.*, 1999) provide statistics for the most common cyclic hydrogen-bonded motifs, and it is likely that motif information will be included in the developing IsoStar knowledge-based library described in Section 22.5.5.8.

22.5.5.7. *The answer 'no'*

Previous sections have illustrated the location and characterization of some important non-covalent interactions. Equally important is a knowledge of when such interactions *do not* occur although chemical sensibility might indicate that they should. We provide four examples from the CSD: (*a*) only 4.8% of more than 1000 thioether S atoms form hydrogen-atom contacts that are within van der Waals limits, despite the obvious analogy with the potent acceptor C—O—C (Allen, Bird *et al.*, 1997*b*); (*b*) of 118 instances in which a furan ring coexists with N—H or O—H donors, the O atom forms hydrogen bonds on only three occasions (Nobeli *et al.*, 1997); (*c*) the ester oxygen $(R_1)(\mathrm{O}{=})\mathrm{C}{-}\underline{\mathrm{O}}{-}R_2$ almost never forms strong hydrogen bonds, although the adjunct carbonyl oxygen atom is well known as a highly potent acceptor (Lommerse *et al.*, 1997); and (*d*) covalently bound fluorine atoms rarely form hydrogen bonds (Dunitz & Taylor, 1997).

22.5.5.8. *IsoStar: a library of non-bonded interactions*

The previous sections show that the amount of data in the CSD on intermolecular geometries is vast, and CSD-derived information for a number of specific systems is available in the literature at various levels of detail. If not, the CSD must be searched for contacts between the relevant functional groups. To provide structured and direct access to a more comprehensive set of derived information, a knowledge-based library of non-bonded interactions (IsoStar: Bruno *et al.*, 1997) has been developed at the CCDC since 1995. IsoStar is based on experimental data, not only from the CSD but also from the PDB, and contains some theoretical results calculated using the IMPT method. Version 1.1 of IsoStar, released in October 1998, contains information on non-bonded interactions formed between 310 common functional groups, referred to as *central groups*, and 45 *contact groups*, *e.g.* hydrogen-bond donors, water, halide ions *etc.* Information is displayed in the form of scatter plots for each interaction. Version 1.1 contains about 12 000 scatter plots: 9000 from the CSD and 3000 from the PDB. IsoStar also reports results for 867 theoretical potential-energy minima.

For a given contact between between a central group (A) and a contact group (B), CSD search results were transformed into an easily visualized form by overlaying the A moieties. This results in a 3D distribution (scatter plot) showing the experimental distribution of B around A. Fig. 22.5.5.1(*a*) shows an example of a scatter plot: the distribution of OH groups around carboxylate anions, illustrating hydrogen-bond formation along the lone-pair directions of the carboxylate oxygens. The IsoStar software provides a tool that enables the user to inspect quickly the original crystal structures in which the contacts occur *via* a hyperlink to the original CSD entries. This is very helpful in identifying outliers, motifs and biases. Another tool generates contoured surfaces from scatter plots, which show the density distribution of the contact groups. A similar approach was first used by Rosenfield *et al.* (1984). Contouring aids the interpretation of the scatter plot and the analysis of preferred geometries. Fig. 22.5.5.1(*b*) shows the contoured surface of the scatter plot in Fig. 22.5.5.1(*a*); the lone-pair directionality now becomes even more obvious.

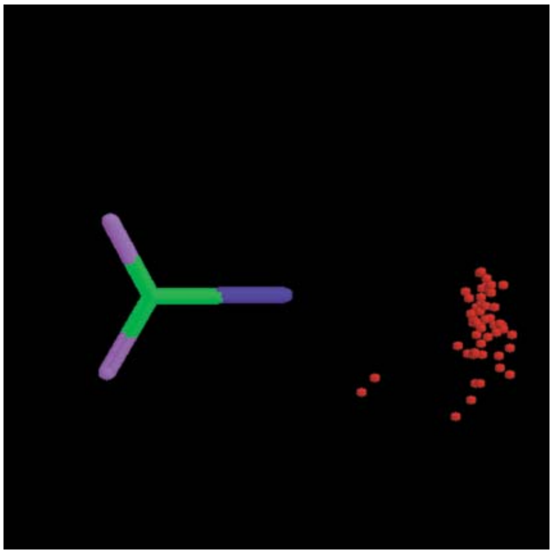


Figure 22.5.5.3
Distribution of oxygen atoms around C(aromatic)—I (CSD data from the IsoStar library, see text).

The fact that carboxylate anions form hydrogen bonds along their lone-pair directions may be well known, although force fields do not always use this information. However, the IsoStar library also contains information on many less well understood functional groups. The interaction between aromatic halo groups and oxygen atoms (Lommerse *et al.*, 1996) is referred to above, and Fig. 22.5.5.3 shows the distribution of oxygen acceptor atoms around aromatic iodine groups. It is clear that the contact O atoms are preferentially observed along the elongation of the C—I bond.

The PDB scatter plots in IsoStar only involve interactions between non-covalently bound ligands and proteins, *i.e* side chain–side chain interactions are excluded. Similar work was presented by Tintelnot & Andrews (1989), but at that time the PDB contained only 40 structures of protein–ligand complexes. The IsoStar library contains data derived from almost 800 complexes having a resolution better than 2.5 Å. Fig. 22.5.5.1(*c*) shows an example of a scatter plot from the PDB (the distribution of OH groups around carboxylate groups). Here, although the hydrogen atoms are missing in the PDB plot, the close similarity between Figs. 22.5.5.1(*c*) (PDB) and 22.5.5.1(*a*) (CSD) is obvious.

22.5.5.9. *Protein–ligand binding*

The reluctance to use data from the CSD because they do not relate directly to biological systems has been noted earlier. However, in principle, the same forces that drive the inclusion of a new molecule into a growing crystal should also apply to the

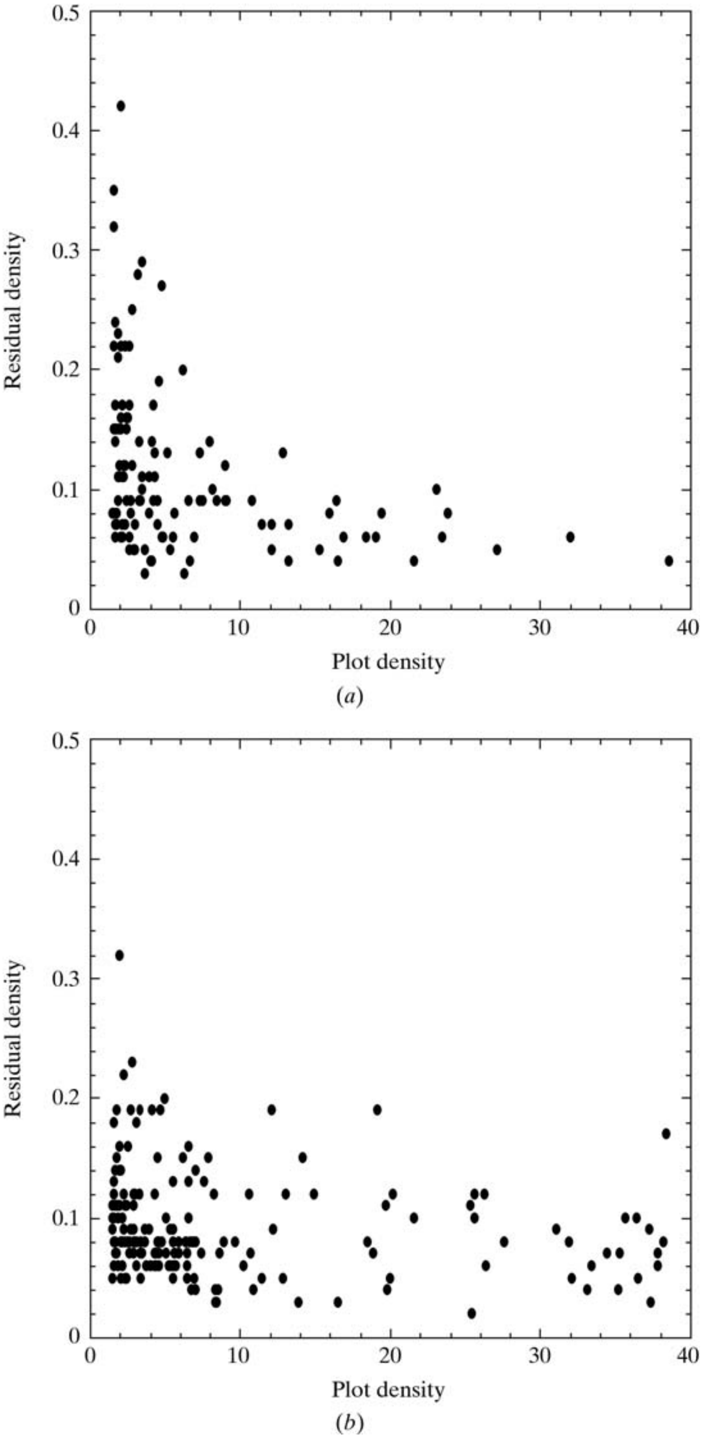


Figure 22.5.5.4
Pairwise comparison of intermolecular-interaction density maps from the CSD and the PDB. Plots of *residual density* $|\rho(\text{CSD}) - \rho(\text{PDB})|$ *versus plot density*, *i.e.* the average density in the least dense situation (CSD or PDB), for situations where the protonation state of the central group is (*a*) unambiguous, and (*b*) ambiguous.

binding of a ligand to a protein. In both cases, molecule and target need to be de-solvated first (although in the first case not necessarily from a water environment) and then interact in the most favourable way.

Nicklaus and colleagues suggested that on average, the conformational energy of ligands in the protein-bound state is 66 (48) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo* (Nicklaus *et al.*, 1995). This result was based on 33 protein–ligand complexes from the PDB for which the ligand also occurs in a small-molecule structure in the CSD. The same investigation also showed that, although ligand conformations in the protein-bound state are generally different from those observed in small-molecule crystal structures, on average the conformational energy of the ligand in the CSD crystal-structure conformation is 66 (47) kJ mol^{-1} above that of the global minimum-energy conformation *in vacuo*, although Boström *et al.* (1998) have shown that these conformational energies are much lower if calculated in a water environment. The computational work indicates that the forces that affect the conformation of a ligand are of comparable magnitude at a protein binding site to those in a small-molecule crystal-structure environment. Thus, if small-molecule crystal-structure statistics tell us that a given structure fragment can only adopt one conformation, generally there is no reason to believe that a ligand that contains this fragment will adopt a different conformation when it binds to a protein.

In principle, the information on non-bonded interactions derived from the CSD and assembled in the IsoStar library should be very important for the understanding and prediction of interaction geometries. However, in light of the comments above, it is important to know whether these data *are* generally relevant to interactions that occur in the protein binding site. Work by Klebe (1994) indicated that, at least for a limited set of test cases, the geometrical distributions derived from ligand–protein complexes are similar to those derived from small-molecule crystal structures. Since the IsoStar library contains information from both the PDB and the CSD, it provides the ultimate basis for establishing similarities (or not) between the interaction geometries observed in small-molecule crystal structures and those observed in protein–ligand complexes. Comparing CSD scatter plots with their corresponding plots from the PDB is an obvious way of establishing the relevance of non-bonded interaction data from small-molecule crystal structures to biological systems.

A full systematic comparison of PDB and CSD scatter plots or, more accurately, of PDB and CSD *density maps* has recently been performed by Verdonk (1998). He calculated residual densities, obtained by subtracting one density map from the other, for each pair of density maps. It appears that, in general, CSD and PDB plots (and thus interaction geometries) are very similar indeed: the average residual density is only 10 (10)%, indicating that 90% of the density in the PDB map is also observed in the CSD map. In Fig. 22.5.5.4(*a*), the average residual densities of each PDB–CSD comparison are plotted *versus* the average concentration of contact groups in the scatter plot. The filled circles represent comparisons for which the protonation state of the central group is unambiguous (*i.e.* carboxylic acid, imidazole *etc.* were excluded). It appears that the residual density decreases with the amount of data in the plots, obviously caused by the more accurate calculation of the residual density. The 'true' residual density seems to be as low as about 6%.

Fig. 22.5.5.4(*b*) shows a similar graph, but now for those density maps in which the protonation state of the central group is ambiguous. As expected, the spread in the calculated residual densities is much higher, even for very dense plots. By comparing the density map from the PDB with the CSD maps for the different protonation states of the central group, the most frequent protonation state of this central group in the protein structures can be predicted. In Table 22.5.5.1, for example, the residual densities for protein carboxylic acid groups are shown, compared with the CSD plots of the neutral carboxylic acid and with those of the charged carboxylate anion. In all cases, the residual density is lower if the PDB map is compared with the CSD map for charged carboxylate anions. This indicates that the

Table 22.5.5.1
Residual densities for carboxylic acid groups

The PDB density maps are compared with the CSD maps for uncharged carboxylic acid and for charged carboxylate anions.

	Residual density (CCO_2H)	Residual density ($CCOO^-$)
Any (N,O,S)—H	0.06	0.04
Any N—H nitrogen	0.07	0.05
Any O—H oxygen	0.07	0.05
Non-donating oxygen	0.12	0.04
Carbonyl oxygen	0.13	0.07
Carbonyl carbon	0.12	0.04
Water oxygen	0.07	0.05
Any aliphatic C—H carbon	0.08	0.06

majority of glutamate and aspartate side chains are charged, which is consistent with other evidence.

22.5.5.10. Modelling applications that use CSD data

Predicting binding modes of ligands at protein binding sites is a problem of paramount importance in drug design. One approach to this problem is to attempt to dock the ligand directly into the binding site. There are several protein–ligand docking programs available, *e.g. DOCK* (see Kuntz *et al.*, 1994), *GRID* (Goodford, 1985), *FLExX* and *FLExS* (Rarey *et al.*, 1996; Lemmen & Lengauer, 1997), and *GOLD* (Jones *et al.*, 1995, 1997). The docking program *GOLD*, developed by the University of Sheffield, Glaxo Wellcome and the CCDC, and which has the high docking success rate of 73%, uses a small torsion library, based on the data from the CSD, to explore the conformational space of the ligand. Its hydrogen-bond geometries and fitness functions are also partly based on CSD data. In the future, we intend to create a more direct link between the crystallographic data and the docking program, *via* IsoStar and the developing torsion library.

Another approach to the prediction of binding modes is to calculate the energy fields for different probes at each position of the binding site, for instance using the *GRID* program (Goodford, 1985). The resulting maps can be displayed as contoured surfaces which can assist in the prediction and understanding of binding modes of ligands. CCDC is developing a program called *SuperStar* (Verdonk *et al.*, 1999) which uses a similar approach to that of the *X-SITE* program (Singh *et al.*, 1991; Laskowski *et al.*, 1996). However, *SuperStar* uses non-bonded interaction data from the CSD rather than the protein side chain–side chain interaction data employed in *X-SITE*. Thus, for a given binding site and contact group (probe), *SuperStar* selects the appropriate scatter plots from the IsoStar library, superimposes the scatter plots on the relevant functional groups in the binding site, and transforms them into one composite probability map. Such maps can then, for example, be used to predict where certain functional groups are likely to interact with the binding site. The strength of *SuperStar* is that it is based entirely on experimental data (although this is also the cause of some limitations). The fields simply represent what has been observed in crystal strucures. We are currently verifying *SuperStar* on a test set of more than 100 protein–ligand complexes from the PDB and preliminary results are encouraging.

Finally, CSD data are used in several *de novo* design programs. These types of programs, *e.g. LUDI* (Böhm, 1992*a*,*b*), predict novel ligands that will interact favourably with a given protein and use hydrogen-bond geometries from the CSD (indirectly) to position their structural fragments in the binding site.

22.5.6. Conclusion

This chapter has summarized the vast range of structural knowledge that can be derived from the small-molecule data contained in the CSD. We have attempted to show that much of this knowledge is directly transferable and applicable to the protein environment. Far from being discrete, structural studies of small molecules and proteins have a natural synergy which, if exploited creatively, will lead to significant advances in both areas. It is therefore unsurprising that some of these CSD studies have been prompted by initial observations made on proteins.

As a result of this activity, it is now very clear that software access to the information stored in the CSD and the PDB must be at two levels: a raw-data level and a derived-knowledge level. The onward development of structural knowledge bases from the underlying data provides for the preservation and storage of the results of data-mining experiments, thus avoiding repetition of standard experiments and providing instant access to complex derivative information. Most importantly, a suitably structured knowledge base can be acted on by software tools that are designed to solve complex problems in structural chemistry (see *e.g.* Thornton & Gardner, 1989; Allen *et al.*, 1990; Bruno *et al.*, 1997; Jones *et al.*, 1997). The availability of knowledge bases derived from experimental observations is likely to be a crucial factor in the solution of those two analogous, and currently intractable, problems in the small-molecule and protein-structure domains: crystal structure and polymorph prediction on the one hand, and protein folding on the other.

References

Abola, E. E., Sussman, J. L., Prilusky, J. & Manning, N. O. (1997). *Protein Data Bank archives of three-dimensional macromolecular structures. Methods Enzymol.* **277**, 556–571.

Allen, F. H., Baalham, C. A., Lommerse, J. P. M. & Raithby, P. R. (1998). *Carbonyl–carbonyl interactions can be competitive with hydrogen bonds. Acta Cryst.* B**54**, 320–329.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*a*). *Resonance-induced hydrogen bonding at sulfur acceptors in* $R_1R_2C{=}S$ *and* $R_1CS_2^-$ *systems. Acta Cryst.* B**53**, 680–695.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997*b*). *Hydrogen-bond acceptor and donor properties of divalent sulfur. Acta Cryst.* B**53**, 696–701.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database system. J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H., Doyle, M. J. & Auf der Heyde, T. P. E. (1991). *Automated conformational analysis from crystallographic data. 6. Principal-component analysis for n-membered carbocyclic rings (n = 4, 5, 6): symmetry considerations and correlations with ring-puckering parameters. Acta Cryst.* B**47**, 412–424.

Allen, F. H., Doyle, M. J. & Taylor, R. (1991). *Automated conformational analysis from crystallographic data. 3. Three-dimensional pattern recognition within the Cambridge Structural Database system: implementation and practical examples. Acta Cryst.* B**47**, 50–61.

Allen, F. H., Harris, S. E. & Taylor, R. (1996). *Comparison of conformer distributions in the crystalline state with conformational energies calculated by ab initio techniques. J. Comput.-Aided Mol. Des.* **10**, 247–254.

Allen, F. H., Howard, J. A. K. & Pitchford, N. A. (1996). *Symmetry-modified conformational mapping and classification of the medium rings from crystallographic data. IV. Cyclooctane and related eight-membered rings. Acta Cryst.* B**52**, 882–891.

Allen, F. H., Kennard, O., Watson, D. G., Brammer, L., Orpen, A. G. & Taylor, R. (1987). *Tables of bond lengths determined by X-ray and neutron diffraction. Part 1. Bond lengths in organic compounds. J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

Allen, F. H., Lommerse, J. P. M., Hoy, V. J., Howard, J. A. K. & Desiraju, G. R. (1997). *Halogen···O(nitro) supramolecular synthon in crystal* *engineering: a combined crystallographic database and ab initio molecular orbital study*. *Acta Cryst.* B**53**, 1006–1016.

Allen, F. H., Motherwell, W. D. S., Raithby, P. R., Shields, G. P. & Taylor, R. (1999). *Systematic analysis of the probabilities of formation of bimolecular hydrogen bonded ring motifs in organic crystal structures*. *New J. Chem.* **23**, 25–34.

Allen, F. H., Raithby, P. R., Shields, G. P. & Taylor, R. (1998). *Probabilities of formation of bimolecular cyclic hydrogen bonded motifs in organic crystal structures: a systematic database study*. *Chem. Commun.* pp. 1043–1044.

Allen, F. H., Rowland, R. S., Fortier, S. & Glasgow, J. I. (1990). *Knowledge acquisition from crystallographic databases: towards a knowledge-based approach to molecular scene analysis*. *Tetrahedron Comput. Methodol.* **3**, 757–774.

Ashida, T., Tsunogae, Y., Tanaka, I. & Yamane, T. (1987). *Peptide chain structure parameters, bond angles and conformation angles from the Cambridge Structural Database*. *Acta Cryst.* B**43**, 212–218.

Balasubramanian, R., Chidambaram, R. & Ramachandran, G. N. (1970). *Potential functions for hydrogen-bond interactions. II. Formulation of an empirical potential function*. *Biochim. Biophys. Acta*, **221**, 196–206.

Berkovitch-Yellin, Z. & Leiserowitz, L. (1984). *The role played by C—H···O and C—H···N interactions in determining molecular packing and conformation*. *Acta Cryst.* B**40**, 159–165.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Bertolasi, V., Gilli, P., Ferretti, V. & Gilli, G. (1996). *Resonance-assisted O—H···O hydrogen bonding. Its role in the crystalline self-recognition of beta-diketone enols and its structural and IR characterisation*. *Chem. Eur. J.* **2**, 925–934.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). *Knowledge-based prediction of protein structures and the design of novel molecules*. *Nature (London)*, **326**, 347–352.

Böhm, H.-J. (1992*a*). *The computer program LUDI: a new method for the de novo design of enzyme inhibitors*. *J. Comput.-Aided Mol. Des.* **6**, 61–78.

Böhm, H.-J. (1992*b*). *LUDI: rule-based automatic design of new substituents for enzyme inhibitors*. *J. Comput.-Aided Mol. Des.* **6**, 593–606.

Bondi, A. (1964). *van der Waals volumes and radii*. *J. Phys. Chem.* **68**, 441–452.

Boström, J., Norrby, P.-O. & Liljefors, T. (1998). *Conformational energy penalties of protein-bound ligands*. *J. Comput.-Aided Mol. Des.* **12**, 383–396.

Bower, M., Cohen, F. E. & Dunbrack, R. L. Jr (1997). *Prediction of protein side-chain rotamers from a backbone-dependent rotamer library*. *J. Mol. Biol.* **267**, 1268–1282.

Brammer, L., Zhao, D., Ladipo, F. T. & Braddock-Wilking, J. (1995). *Hydrogen bonds involving transition metal centres – a brief review*. *Acta Cryst.* B**51**, 632–640.

Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *IsoStar: a library of information about nonbonded interactions*. *J. Comput.-Aided Mol. Des.* **11**, 525–537.

Bürgi, H.-B. & Dunitz, J. D. (1983). *From crystal statics to chemical dynamics*. *Acc. Chem. Res.* **16**, 153–161.

Bürgi, H.-B. & Dunitz, J. D. (1988). *Can statistical analysis of structural parameters from different crystal environments lead to quantitative energy relationships?* *Acta Cryst.* B**44**, 445–451.

Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure Correlation*. Weinheim: VCH Publishers.

Carbonell, J. (1989). Editor. *Machine Learning – Paradigms and Methods*. Amsterdam: Elsevier.

Carrell, A. B., Shimoni, L., Carrell, C. J., Bock, C. W., Murray-Rust, P. & Glusker, J. P. (1993). *The stereochemistry of the recognition of nitrogen-containing heterocycles by hydrogen bonding and by metal ions*. *Receptor*, **3**, 57–76.

Carrell, C. J., Carrell, H. L., Erlebacher, J. & Glusker, J. P. (1988). *Structural aspects of metal ion–carboxylate interactions*. *J. Am. Chem. Soc.* **110**, 8651–8656.

Ceccarelli, C., Jeffrey, G. A. & Taylor, R. (1981). *A survey of O—H···O hydrogen bond geometries determined by neutron diffraction*. *J. Mol. Struct.* **70**, 255–271.

Chakrabarti, P. (1990*a*). *Interaction of metal ions with carboxylic and carboxamide groups in protein structures*. *Protein. Eng.* **4**, 49–56.

Chakrabarti, P. (1990*b*). *Geometry of interaction of metal ions with histidine residues in protein structures*. *Protein Eng.* **4**, 57–63.

Chakrabarti, P. & Dunitz, J. D. (1982). *Directional preferences of ether O-atoms towards alkali and alkaline earth cations*. *Helv. Chim. Acta*, **65**, 1482–1488.

Chatfield, C. & Collins, A. J. (1980). *Introduction to Multivariate Analysis*. London: Chapman & Hall.

Conklin, D., Fortier, S., Glasgow, J. I. & Allen, F. H. (1996). *Conformational analysis from crystallographic data using conceptual clustering*. *Acta Cryst.* B**52**, 535–549.

Cremer, D. & Pople, J. A. (1975). *A general definition of ring puckering coordinates*. *J. Am. Chem. Soc.* **97**, 1354–1358.

Deane, C. M., Allen, F. H., Taylor, R. & Blundell, T. L. (1999). *Carbonyl–carbonyl interactions stabilise the partially allowed Ramachandran conformations of asparagine and aspartic acid*. *Protein Eng.* **12**, 1025–1028.

Derewenda, Z. S., Lee, L. & Derewenda, U. (1995). *The occurrence of C—H···O hydrogen bonds in proteins*. *J. Mol. Biol.* **252**, 248–262.

Desiraju, G. R. (1989). *Crystal Engineering: The Design of Organic Solids*. New York: Academic Press.

Desiraju, G. R. (1991). *The C—H···O hydrogen bond in crystals. What is it?* *Acc. Chem. Res.* **24**, 270–276.

Desiraju, G. R. (1995). *Supramolecular synthons in crystal engineering – a new organic synthesis*. *Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.

Desiraju, G. R., Kashino, S., Coombs, M. M. & Glusker, J. P. (1993). *C—H···O packing motifs in some cyclopenta[a]phenanthrenes*. *Acta Cryst.* B**49**, 880–892.

Desiraju, G. R. & Murty, B. N. (1987). *Correlation between crystallographic and spectroscopic properties for C—H···O bonds in terminal acetylenes*. *Chem. Phys. Lett.* **139**, 360–361.

Donohue, J. (1952). *The hydrogen bond in organic crystals*. *J. Phys. Chem.* **56**, 502–510.

Donohue, J. (1968). *Selected topics in hydrogen bonding*. In *Structural Chemistry and Molecular Biology*, edited by W. Davidson & E. Rich, pp. 443–465. San Francisco: W. H. Freeman.

Dunbrack, R. L. Jr & Karplus, M. (1993). *Backbone-dependent rotamer library for proteins: applications to side-chain prediction*. *J. Mol. Biol.* **230**, 534–571.

Dunitz, J. D. & Taylor, R. (1997). *Organic fluorine hardly ever accepts hydrogen bonds*. *Chem. Eur. J.* **3**, 83–90.

Einspahr, H. & Bugg, C. E. (1981). *The geometry of calcium–carboxylate interactions in crystalline complexes*. *Acta Cryst.* B**37**, 1044–1052.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement*. *Acta Cryst.* A**47**, 392–400.

Everitt, B. (1980). *Cluster Analysis*. New York: Wiley.

Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Molecular scene analysis: the integration of direct-methods and artificial-intelligence strategies for solving protein crystal structures*. *Acta Cryst.* D**49**, 168–178.

Glusker, J. P. (1980). *Citrate conformation and chelation: enzymatic implications*. *Acc. Chem. Res.* **13**, 345–352.

Glusker, J. P., Lewis, M. & Rossi, M. (1994). *Crystal Structure Analysis for Chemists and Biologists*. Weinheim: VCH Publishers.

Goodford, P. J. (1985). *A computational procedure for determining energetically favourable binding sites on biologically important molecules*. *J. Med. Chem.* **28**, 849–857.

Gould, R. O., Gray, A. M., Taylor, P. & Walkinshaw, M. D. (1985). *Crystal environments and geometries of leucine, isoleucine, valine and phenylalanine provide estimates of minimum nonbonded contact and preferred van der Waals interaction distances*. *J. Am. Chem. Soc.* **107**, 5921–5927.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography*. *Acta Cryst.* A**47**, 655–685.

Hayes, I. C. & Stone, A. J. (1984). *An intermolecular perturbation theory for the region of moderate overlap*. *J. Mol. Phys.* **53**, 83–105.

Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Errors in protein structures*. *Nature (London)*, **381**, 272.

Jeffrey, G. A. & Maluszynska, H. (1982). *A survey of hydrogen bond geometries in the crystal structures of amino acids. Int. J. Biol. Macromol.* **4**, 173–185.

Jeffrey, G. A. & Maluszynska, H. (1986). *A survey of the geometry of hydrogen bonds in the crystal structures of barbiturates, purines and pyrimidines. J. Mol. Struct.* **147**, 127–142.

Jeffrey, G. A. & Mitra, J. (1984). *Three-centre (bifurcated) hydrogen bonding in the crystal structures of amino acids. J. Am. Chem. Soc.* **106**, 5546–5553.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures*. Berlin: Springer Verlag.

Jones, G., Willett, P. & Glen, R. C. (1995). *Molecular recognition of receptor sites using a genetic algorithm with a description of solvation. J. Mol. Biol.* **245**, 43–53.

Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. (1997). *Development and validation of a genetic algorithm for flexible docking. J. Mol. Biol.* **267**, 727–748.

Joris, L., Schleyer, P. & Gleiter, R. (1968). *Cyclopropane rings as proton acceptor groups in hydrogen bonding. J. Am. Chem. Soc.* **90**, 327–336.

Kennard, O. (1962). In *International Tables for X-ray Crystallography*, Vol. II. Birmingham: Kynoch Press.

Kennard, O. & Allen, F. H. (1993). *The Cambridge Crystallographic Data Centre. Chem. Des. Autom. News*, **8**, 1, 31–37.

Klebe, G. (1994). *The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands. J. Mol. Biol.* **237**, 212–235.

Klebe, G. & Mietzner, T. (1994). *A fast and efficient method to generate biologically relevant conformations. J. Comput.-Aided Mol. Des.* **8**, 583–594.

Kroon, J. & Kanters, J. A. (1974). *Non-linearity of hydrogen bonds in molecular crystals. Nature (London)*, **248**, 667–669.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vliegenthart, J. A. (1975). *O—H···O hydrogen bonds in molecular crystals: a statistical and quantum chemical analysis. J. Mol. Struct.* **24**, 109–129.

Kuntz, I. D., Meng, E. C. & Stoichet, B. K. (1994). *Structure-based molecular design. Acc. Chem. Res.* **27**, 117–122.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst.* **26**, 283–291.

Laskowski, R. A., Thornton, J. M., Humblet, C. & Singh, J. (1996). *X-SITE: use of empirically derived atomic packing preferences to identify favourable interaction regions in the binding sites of proteins. J. Mol. Biol.* **259**, 175–201.

Lehn, J.-M. (1988). *Perspectives in supramolecular chemistry – from molecular recognition towards molecular information processing and self-organization. Angew. Chem. Int. Ed. Engl.* **27**, 90–112.

Lemmen, C. & Lengauer, T. (1997). *Time-efficient flexible superposition of medium sized molecules. J. Comput.-Aided Mol. Des.* **11**, 357–368.

Levitt, M. & Perutz, M. (1988). *Aromatic rings act as hydrogen bond acceptors. J. Mol. Biol.* **201**, 751–754.

Lommerse, J. P. M., Price, S. L. & Taylor, R. (1997). *Hydrogen bonding of carbonyl, ether and ester oxygen atoms with alkanol hydroxyl groups. J. Comput. Chem.* **18**, 757–780.

Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). *The nature and geometry of intermolecular interactions between halogens and oxygen or nitrogen. J. Am. Chem. Soc.* **118**, 3108–3116.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*a*). *Coulombic interactions between partially charged main-chain atoms not hydrogen bonded to each other influence the conformations of α-helices and antiparallel β-sheets. J. Mol. Biol.* **248**, 361–373.

Maccallum, P. H., Poet, R. & Milner-White, E. J. (1995*b*). *Coulombic interactions between partially charged main-chain atoms stabilise the right-handed twist found in most β-strands. J. Mol. Biol.* **248**, 374–384.

Miller, J., McLachlan, A. D. & Klug, A. (1985). *Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J.* **4**, 1609–1614.

Murray-Rust, P. & Bland, R. (1978). *Computer retrieval and analysis of molecular geometry. II. Variance and its interpretation. Acta Cryst.* B**34**, 2527–2533.

Murray-Rust, P. & Glusker, J. P. (1984). *Directional hydrogen bonding to sp^2 and sp^3 hybridized oxygen atoms and its relevance to ligand–macromolecule interactions. J. Am. Chem. Soc.* **105**, 1018–1025.

Murray-Rust, P. & Motherwell, S. (1978). *Computer retrieval and analysis of molecular geometry. III. Geometry of the β-1′-aminofuranoside fragment. Acta Cryst.* B**34**, 2534–2546.

Nicklaus, M. C., Wang, S., Driscoll, J. S. & Milne, G. W. A. (1995). *Conformational changes of small molecules binding to proteins. Bioorg. Med. Chem.* **3**, 411–428.

Nobeli, I., Price, S. L., Lommerse, J. P. M. & Taylor, R. (1997). *Hydrogen bonding properties of oxygen and nitrogen acceptors in aromatic heterocycles. J. Comput. Chem.* **18**, 2060–2074.

Nyburg, S. C. & Faerman, C. H. (1985). *A revision of van der Waals atomic radii for molecular crystals: N, O, F, S, Cl, Se, Br and I bonded to carbon. Acta Cryst.* B**41**, 274–279.

Orpen, A. G., Brammer, L., Allen, F. H., Kennard, O., Watson, D. G. & Taylor, R. (1989). *Tables of bond lengths determined by X-ray and neutron diffraction. Part II: organometallic compounds and coordination complexes of the d- and f-block metals. J. Chem. Soc. Dalton Trans.* pp. S1–S83.

Pauling, L. (1939). *The Nature of the Chemical Bond*. Ithaca: Cornell University Press.

Pedireddi, V. R. & Desiraju, G. R. (1992). *A crystallographic scale of carbon acidity. Chem. Commun.* pp. 988–990.

Pimentel, G. C. & McClellan, A. L. (1960). *The Hydrogen Bond*. San Francisco: W. H. Freeman.

Price, S. L., Stone, A. J., Lucas, J., Rowland, R. S. & Thornley, A. E. (1994). *The nature of —Cl···Cl— intermolecular interactions. J. Am. Chem. Soc.* **116**, 4910–4918.

Rappoport, Z., Biali, S. E. & Kaftory, M. (1990). *Application of the structure correlation method to ring-flip processes in benzophenone. J. Am. Chem. Soc.* **112**, 7742–7750.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *Predicting receptor-ligand interactions by an incremental construction algorithm. J. Mol. Biol.* **261**, 470–481.

Robertson, J. M. (1953). *Organic Crystals and Molecules*. Ithaca: Cornell University Press.

Rosenfield, R. E., Parthasarathy, R. & Dunitz, J. D. (1977). *Directional preferences of non-bonded atomic contacts with divalent sulphur. 1. Electrophiles and nucleophiles. J. Am. Chem. Soc.* **99**, 4860–4862.

Rosenfield, R. E. Jr, Swanson, S. M., Meyer, E. F. Jr, Carrell, H. L. & Murray-Rust, P. (1984). *Mapping the atomic environment of functional groups: turning 3D scatterplots into pseudo-density contours. J. Mol. Graphics*, **2**, 43–46.

Rowland, R. S. & Taylor, R. (1996). *Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii. J. Phys. Chem.* **100**, 7384–7391.

Schweizer, W. B. & Dunitz, J. D. (1982). *Structural characteristics of the carboxylic acid ester group. Helv. Chim. Acta*, **65**, 1547–1552.

Singh, J., Saldanha, J. & Thornton, J. M. (1991). *A novel method for the modelling of peptide ligands to their receptors. Protein Eng.* **4**, 251–261.

Steiner, T., Kanters, J. A. & Kroon, J. (1996). *Acceptor directionality of sterically unhindered C—H···O═C hydrogen bonds donated by acidic C—H groups. J. Chem. Soc. Chem. Commun.* **11**, 1277–1278.

Steiner, T. & Saenger, W. (1992). *Geometry of C—H···O hydrogen bonds in carbohydrate crystal structures. Analysis of neutron diffraction data. J. Am. Chem. Soc.* **114**, 10146–10154.

Steiner, T., Starikov, E. B., Amado, A. M. & Teixeira-Dias, J. J. C. (1995). *Weak hydrogen bonding. Part 2. The hydrogen-bonding nature of short C—H··· · contacts. Crystallographic, spectroscopic and quantum mechanistic studies of some terminal alkynes. J. Chem. Soc. Perkin Trans.* 2, pp. 1312–1326.

Strynadka, N. C. J. & James, M. N. G. (1989). *Crystal structures of the helix-loop-helix calcium-binding proteins. Annu. Rev. Biochem.* **58**, 951–998.

Sutton, L. E. (1956). *Tables of Interatomic Distances and Configuration in Molecules and Ions*. Special Publication No. 11. London: The Chemical Society.

Sutton, L. E. (1959). *Tables of Interatomic Distances and Configuration in Molecules and Ions (Supplement)*. Special Publication No. 18. London: The Chemical Society.

Taylor, R. (1986). *The Cambridge Structural Database in molecular graphics: techniques for the rapid identification of conformational minima. J. Mol. Graphics*, **4**, 123–131.

Taylor, R. & Allen, F. H. (1994). *Statistical and numerical methods of data analysis*. In *Structure Correlation*, edited by H.-B. Bürgi & J. D. Dunitz. Weinheim: VCH Publishers.

Taylor, R. & Kennard, O. (1982). *Crystallographic evidence for the existence of C—H···O, C—H···N and C—H···Cl hydrogen bonds. J. Am. Chem. Soc.* **104**, 5063–5070.

Taylor, R. & Kennard, O. (1983). *Comparison of X-ray and neutron diffraction results for the N—H···O═C hydrogen bond. Acta Cryst.* B**39**, 133–138.

Taylor, R., Kennard, O. & Versichel, W. (1983). *Geometry of the N—H···O═C hydrogen bond. 1. Lone-pair directionality. J. Am. Chem. Soc.* **105**, 5761–5766.

Taylor, R., Kennard, O. & Versichel, W. (1984*a*). *Geometry of the N—H···O═C hydrogen bond. 2. Three-centre (bifurcated) and four-centre (trifurcated) bonds. J. Am. Chem. Soc.* **106**, 244–248.

Taylor, R., Kennard, O. & Versichel, W. (1984*b*). *Geometry of the N—H···O═C hydrogen bond. 3. Hydrogen-bond distances and angles. Acta Cryst.* B**40**, 280–288.

Taylor, R., Mullaley, A. & Mullier, G. W. (1990). *Use of crystallographic data in searching for isosteric replacements: composite field environments of nitro and carbonyl groups. Pestic. Sci.* **29**, 197–213.

Thornton, J. M. & Gardner, S. P. (1989). *Protein motifs and database searching. Trends Biochem. Sci.* **14**, 300–304.

Tintelnot, M. & Andrews, P. (1989). *Geometries of functional group interactions in enzyme–ligand complexes: guides for receptor modelling. J. Comput.-Aided Mol. Des.* **3**, 67–84.

Verdonk, M. L. (1998). Unpublished results.

Verdonk, M. L., Cole, J. C. & Taylor, R. (1999). *SuperStar: a knowledge-based approach for identifying interaction sites in proteins. J. Mol. Biol.* **289**, 1093–1108.

Viswamitra, M. A., Radhakrishnan, R., Bandekar, J. & Desiraju, G. R. (1993). *Evidence for O—H···C and N—H···C hydrogen bonding. J. Am. Chem. Soc.* **115**, 4868–4869.

Whitesides, G. M., Simanek, E. E., Mathias, J. P., Seto, C. T., Chin, D. N., Mammen, M. & Gordon, D. M. (1995). *Non-covalent synthesis: using physical-organic chemistry to make aggregates. Acc. Chem. Res.* **28**, 37–43.

International Tables for Crystallography (2012). Vol. F, Chapter 23.1, pp. 749–751.

PART 23. STRUCTURAL ANALYSIS AND CLASSIFICATION

Chapter 23.1. Protein-fold classification

C. Orengo and J. Thornton

Since the first structure of myoglobin was solved in 1971, there has been an exponential growth in known protein structures with about 73 000 structures currently deposited in the Protein Data Bank (PDB; Abola *et al.*, 1987) and 500 or more solved each month. Structural genomics (SG) initiatives are now contributing significantly to this total; between 2005 and 2008 they accounted for some 18% of new structures, and, now that high-throughput pipelines have been set up, some 600–800 structures deposited to the PDB per year (around two structures a day) are from SG centres around the world (Rajesh *et al.*, 2009). When dealing with such large numbers it is necessary to organize the data in a manageable and biologically meaningful way. To this end, several structural classifications have been developed [SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997), DALI (Holm & Sander, 1999), 3Dee (Barton, 1997), HOMSTRAD (Mizuguchi *et al.*, 1998) and ENTREZ (Hogue *et al.*, 1996)], differing in their methodology and the degree of structural and functional annotation for the protein families identified.

Most public classification schemes have chosen to group proteins according to similarities in their domain structures, as this is generally considered to be the important evolutionary and folding unit. However, it can be difficult to identify domain boundaries either manually or using automatic algorithms, and although there are many methods available, it has been shown that even the most reliable algorithms only give the correct answer about 80% of the time (Jones *et al.*, 1998). Methods for recognizing domains are described in Chapter 23.2.

Most protocols used for clustering protein domain structures into families first identify similarities in their sequences. There are many well established methods for doing this, most based on dynamic programming algorithms, and since proteins with sequence identities of 30% or more are known to adopt very similar folds (Sander & Schneider, 1991; Flores *et al.*, 1993), it is relatively simple to cluster related proteins into evolutionary families on this basis. Very distant relatives (<20% sequence identity) are not easily identified by sequence alignment, but since structure is much more highly conserved during evolution, these relationships can be detected by comparing the 3D structures directly.

Various powerful algorithms have been developed for recognizing structurally related proteins (for reviews see Holm & Sander, 1994; Brown *et al.*, 1996). These build on the rigid-body superposition methods of Rossmann & Argos (1975), which compare intermolecular distances after optimal translation and rotation of one protein structure onto the other. Other methods are based on the distance plots developed by Phillips (1970), which enable comparison of intramolecular distances between protein structures. In comparing very distantly related proteins, there are a number of problems which must be overcome. Insertions or deletions can obscure equivalent regions, though generally these appear in the loops between secondary structures. Residue substitutions can cause shifts in the orientations of the secondary structures in order to maintain optimal hydrophobic packing in the core.

A number of strategies have been developed for handling these problems. For example, some methods only consider secondary-structure elements, as these will contain fewer insertions. Artymiuk *et al.* (1989) represent secondary structures as linear vectors and use fast, efficient comparison algorithms based on graph theory. Others have adapted rigid-body methods to optimally superpose secondary structures, ignoring loops. Some methods chop the proteins being compared into fragments and then use various energy-minimization approaches (*e.g.* simulated annealing, Monte Carlo optimization) to link equivalent fragments in the two proteins. Such fragments can be identified by rigid-body superposition (Vriend & Sander, 1991) or, in the case of the DALI method (Holm & Sander, 1994), by comparing contact maps for hexapeptide fragments. Several groups have modified the dynamic programming algorithms designed to cope with insertions or deletions in sequence comparison in order to compare three-dimensional (3D) information (Taylor & Orengo, 1989; Sali & Blundell, 1990; Russell & Barton, 1993). For example, the SSAP method of Taylor & Orengo (1989) uses double dynamic programming to align residue structural environments defined by vectors between Cβ atoms, whilst in STAMP (Russell & Barton, 1993) dynamic programming is used in an iterative procedure, together with rigid-body superposition.

Once equivalent residues have been found, the degree of structural similarity between two proteins can be measured in a number of ways, though the most commonly used is the root-mean-square deviation (RMSD), which is effectively the average 'distance' between superposed residues. However, there is still no consensus about which thresholds might imply homologous proteins or fold similarity between analogous proteins or common structural motifs. It is likely that this will become clearer as more structures are determined and the families become more highly populated, providing more information on tolerance to structural changes. These contraints will probably reflect functional requirements and/or kinetic or thermodynamic factors and will be specific to the family.

Several groups (Holm & Sander, 1999; Hogue *et al.*, 1996) attempt to determine the significance of a structural match by considering the distribution of scores for unrelated proteins and calculating a Z score. These approaches are very reliable for proteins possessing unusual structural characteristics but may not be as sensitive for those with highly recurring and common structural motifs. Other groups use empirical approaches (Orengo *et al.*, 1997) to establish reasonable cutoffs for identifying homologues, though these approaches obviously suffer from the currently limited size of the structure data bank.

Because of the individual strategies used to recognize relatives, the protein-structure classifications differ somewhat in their assignments. However, most classifications group proteins having highly similar sequences ($\geq$30%) into families. Subsequently,

Copyright © 2012 International Union of Crystallography

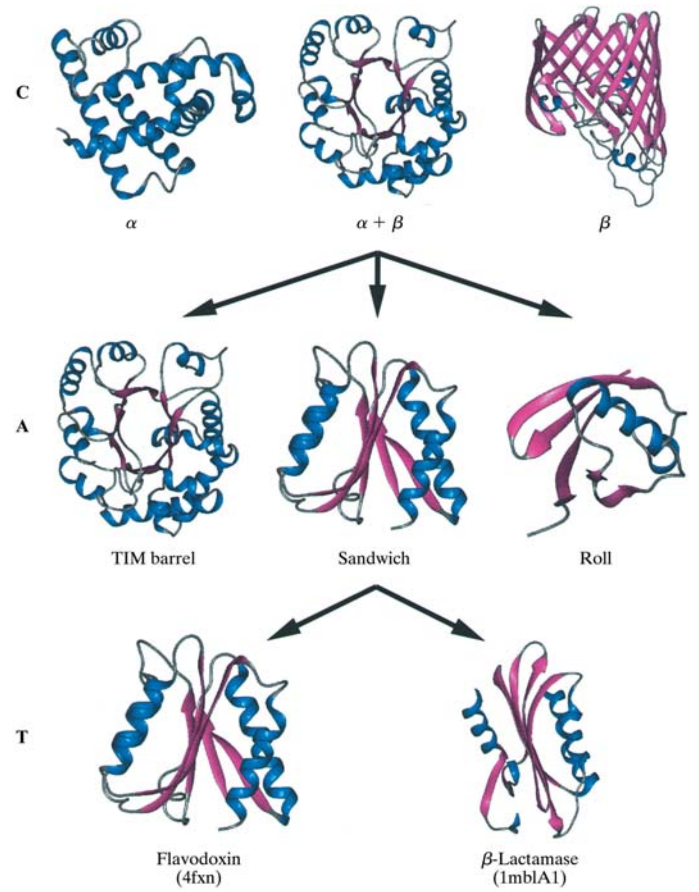


Figure 23.1.1
Schematic representation of the (C)lass, (A)rchitecture and (T)opology/fold levels in the CATH database.

those families having highly similar structures and some other evidence of common ancestry [*e.g.* similar functions or some residual sequence identity (Orengo *et al.*, 1999)] are merged into homologous superfamilies. Families adopting similar folds, but where there is no other evidence to suggest divergent evolution, are usually put into the same fold group but are described as analogous proteins, since their similarity may simply reflect the physical and/or chemical constraints on protein folding.

Recent analysis of CATH superfamilies has shown that some of the most highly populated superfamilies in CATH are structurally very diverse. Typically, domain relatives possess a common 'structural core' but secondary structural embellishments to this core can be considerable and effectively result in homologous structures varying in size by threefold or more and having rather different global folds (Reeves *et al.*, 2006; Greene *et al.*, 2007; Cuff *et al.*, 2009). This poses some challenges to hierarchical domain structure classifications like SCOP and CATH. However, usually the evolutionary conserved cores of relatives within a domain superfamily possess highly similar folds.

SCOP and CATH are currently the largest of the public classifications, each with over 1900 homologous superfamilies. In SCOP (Murzin *et al.*, 1995), these families have been very carefully manually validated using biochemical information and by consideration of special structural features (*e.g.* rare β-bulges, left-handed helical connections) that may constitute evolutionary fingerprints; in CATH, homologues are validated both manually and automatically (Orengo *et al.*, 1997). Other databases [HOMSTRAD (Mizuguchi *et al.*, 1998); 3Dee (Barton, 1997)] contain similar groupings of protein structures, and there are multiple structural alignments for the family, annotated according to residue properties.

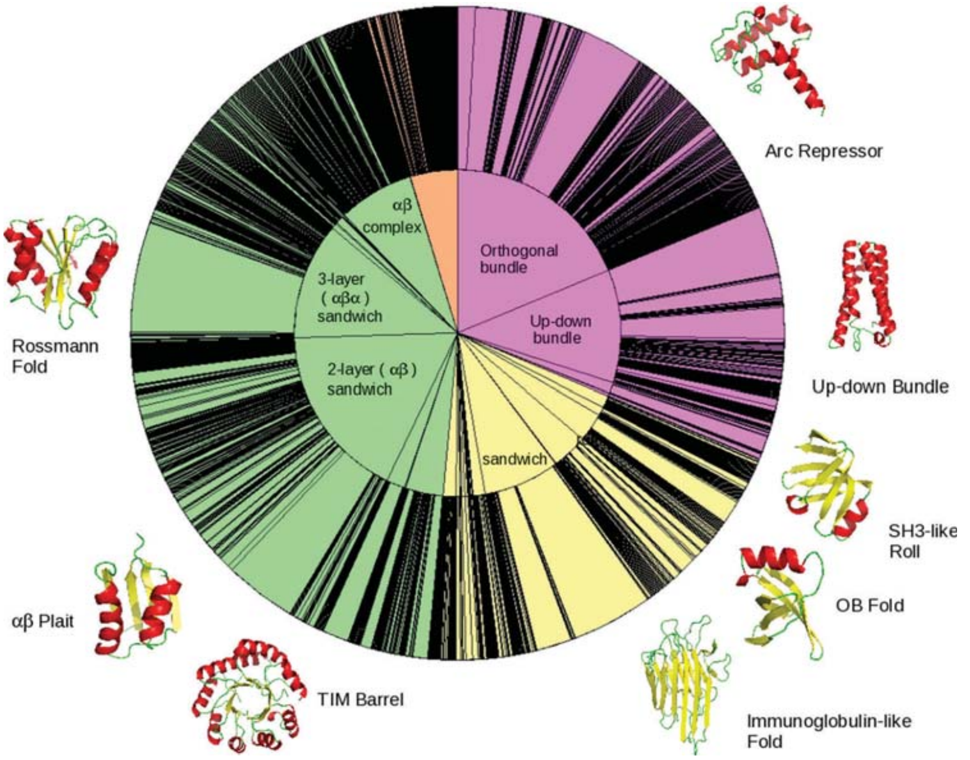


Figure 23.1.2
'CATHerine wheel' plot showing the distribution of non-homologous structures [*i.e.* a single representative from each homologous superfamily (H level) in CATH] amongst the different classes (C), architectures (A) and fold families (T) in the CATH database. Protein classes are shown coloured as red (mainly α), green (mainly β), yellow (α–β) and orange (few secondary structures). Within each class, the angle subtended for a given segment reflects the proportion of structures within the identified architectures (inner circle) or fold groups (outer circle). *MOLSCRIPT* (Kraulis, 1991) illustrations are shown for representative examples from the superfold families.

Several studies have suggested a limited number of folds available to proteins, with estimates ranging from one thousand to several thousand (Chothia, 1993; Orengo *et al.*, 1994), and this will mean an increasing number of analogous protein pairs being identified as the structural genomics initiatives continue. Recent analyses of the population of different fold families have revealed that some folds are more highly populated, perhaps because they fold more easily or are more stable. In the CATH database, ten favoured folds, described as superfolds, comprised very regular, layered architectures and were shown to contain a higher proportion of favoured motifs (*e.g.* Greek key, $\beta\alpha$ motif) than non-superfold structures. Similarly, analysis of SCOP (Brenner *et al.*, 1996) revealed some 40 or so frequently occurring domains (FODS), which included the superfolds. About one-third of all non-homologous structures (<25% sequence identity to each other) adopt one of these folds.

Some groups avoid explicit definition of protein families. The DALI database of Holm & Sander (1999) is a neighbourhood scheme listing all related proteins for a given protein structure.

Neighbours are identified using the DALI structure-comparison algorithm (Holm & Sander, 1993) and range from the most highly similar, homologous proteins to those sharing only motif similarities. The ENTREZ database (Hogue *et al.*, 1996) provides a similar scheme, generated by the VAST structure-comparison method of Gibrat *et al.* (1997). Both allow the user to assess significance and draw their own inferences regarding evolutionary relationships. More recently, the DALI domain database (DDD) (Holm & Sander, 1998) has provided clusters of related proteins based on calculated Z scores.

Most available databases further classify the fold groups on the basis of class. These agree with the major classes recognized by Levitt & Chothia (1976) (mainly α, mainly β, α/β, $\alpha + \beta$), although in the CATH database the α/β and $\alpha + \beta$ classes have been merged (Fig. 23.1.1). CATH also describes an intermediate architecture level between class and fold group (Orengo *et al.*, 1997). This refers to the arrangement of secondary-structure elements in 3D, regardless of their connectivity, and so defines the shape (*e.g.* barrel, sandwich, propeller) (Fig. 23.1.2). There are currently 40 different architectures in CATH, with the simple barrel and sandwich shapes accounting for about 60% of the non-homologous structures.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Protein Data Bank*. In *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132.

Artymiuk, P. J., Mitchell, E. M., Rice, D. W. & Willett, P. (1989). *Searching techniques for databases of protein structures*. *J. Inf. Sci.* **15**, 287–298.

Barton, G. J. (1997). *3Dee: database of protein domain definitions*. http://barton.ebi.ac.uk/servers/3Dee.html.

Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). *Understanding protein structure. Using SCOP for fold interpretation*. *Methods Enzymol.* **266**, 635–643.

Brown, N. P., Orengo, C. A. & Taylor, W. R. (1996). *A protein structure comparison methodology*. *Comput. Chem.* **20**, 359–380.

Chothia, C. (1993). *One thousand families for the molecular biologist*. *Nature (London)*, **357**, 543–544.

Cuff, A. L., Sillitoe, I., Lewis, T., Redfern, O. C., Garratt, R., Thornton, J. & Orengo, C. A. (2009). *The CATH classification revisited – architectures reviewed and new ways to characterize structural divergence in superfamilies*. *Nucleic Acids Res.* **37** (database issue), D310–D314.

Flores, T. P., Orengo, C. A. & Thornton, J. M. (1993). *Conformational characteristics in structurally similar protein pairs*. *Protein Sci.* **7**, 31–37.

Gibrat, J. F., Madej, T., Spouge, J. L. & Bryant, S. H. (1997). *The VAST protein structure comparison method*. *Biophys. J.* **72**, MP298.

Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Peral, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. & Orengo, C. A. (2007). *The CATH domain database: new protocols and classification levels give a more comprehensive resource for exploring evolution*. *Nucleic Acids Res.* **35** (database issue), D291–D297.

Hogue, C. W., Ohkawa, H. & Bryant, S. H. (1996). *A dynamic look at structures: WWW-Entrez and the molecular modelling database*. *Trends Biochem. Sci.* **21**, 226–229.

Holm, L. & Sander, C. (1993). *Protein structure comparison by alignment of distance matrices*. *J. Mol. Biol.* **233**, 123–138.

Holm, L. & Sander, C. (1994). *Searching protein structure databases has come of age*. *Proteins*, **19**, 165–173.

Holm, L. & Sander, C. (1998). *Dictionary of recurrent domains in protein structures*. *Proteins*, **33**, 88–96.

Holm, L. & Sander, C. (1999). *Protein folds and families: sequence and structure alignments*. *Nucleic Acids Res.* **27**, 244–247.

Jones, S., Stewart, M., Michie, A. D., Swindells, M. B., Orengo, C. A. & Thornton, J. M. (1998). *Domain assignment for protein structures using a consensus approach: characterisation and analysis*. *Protein Sci.* **7**, 233–242.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures*. *J. Appl. Cryst.* **24**, 946–950.

Levitt, M. & Chothia, C. (1976). *Structural patterns in globular proteins. Nature (London)*, **261**, 552–558.

Mizuguchi, K., Deane, C. A., Blundell, T. L. & Overerington, J. P. (1998). *HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci.* **7**, 2469–2471.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of the protein database for the investigation of sequences and structures. J. Mol. Biol.* **247**, 536–540.

Orengo, C. A., Jones, D. T., Taylor, W. & Thornton, J. M. (1994). *Protein superfamilies and domain superfolds. Nature (London)*, **372**, 631–634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *CATH – a hierarchic classification of protein domain structures. Structure*, **5**, 1093–1108.

Orengo, C. A., Pearl, F. M. G., Bray, J. E., Todd, A. E., Martin, A. C., LoConte, L. & Thornton, J. M. (1999). *The CATH database provides insights into protein structure/function relationships. Nucleic Acids Res.* **27**, 275–279.

Phillips, D. E. (1970). *British Biochemistry, Past and Present*, p. 11. London Biochemistry Society Symposium. Academic Press.

Rajesh, N., Liu, J., Soong, T., Acton, T. B., Everett, J. K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., Montelione, G. T. & Rost, B. (2009). *Structural genomics is the largest contributor of novel structural leverage. J. Struct. Funct. Genomics*, **10**, 181–191.

Reeves, G., Dallman, T., Redfern, O., Akpor, A. & Orengo, C. A. (2006). *Structural diversity of domain superfamilies in the CATH database. J. Mol. Biol.* **360**, 725–741.

Rossmann, M. G. & Argos, P. (1975). *A comparison of the heme binding pocket in globins and cytochrome b5. J. Biol. Chem.* **250**, 7525–7532.

Russell, R. B. & Barton, G. J. (1993). *Multiple protein sequence alignment from tertiary structure comparisons. Assignments of global and residue level confidences. Proteins*, **14**, 309–323.

Sali, A. & Blundell, T. B. (1990). *The definition of general topological equivalences in proteins: a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J. Mol. Biol.* **212**, 403–428.

Sander, C. & Schneider, R. (1991). *Database of homology-derived protein structures and structural meaning of sequence alignments. Proteins*, **9**, 56–68.

Taylor, W. R. & Orengo, C. A. (1989). *Protein structure alignment. J. Mol. Biol.* **208**, 1–22.

Vriend, G. & Sander, C. (1991). *Detection of common three-dimensional substructures in proteins. Proteins*, **11**, 552–558.

International Tables for Crystallography (2012). Vol. F, Chapter 23.2, pp. 752–754.

Chapter 23.2. Locating domains in three-dimensional structures

L. Holm and C. Sander

23.2.1. Introduction

Modular design is beneficial in many areas of life, including computer programming, manufacturing, and even in protein folding.

Protein-structure analysis has long operated with the notion of domains, *i.e.*, dividing large structures into quasi-independent substructures or modules (Wetlaufer, 1973; Bork, 1992). In various contexts, these substructures are thought to fold autonomously, to carry specific molecular functions such as binding or catalysis, to move relative to each other as semi-rigid bodies and to speed the evolution of new functions by recombination (Fig. 23.2.1.1).

The problem of subdividing protein molecules into structural and functional units has received the attention of numerous researchers over the last 25 years. Early algorithms focused on protein folding or unfolding pathways and aimed at identifying substructures that would be physically stable on their own. Nowadays, with bulging macromolecular databases, the focus has shifted to devise automatic methods for identifying domains that can form the basis for a consistent protein-structure classification (Murzin *et al.*, 1995; Orengo *et al.*, 1997; Holm & Sander, 1999).

This review presents the concepts underlying computational methods for locating domains in three-dimensional structures. Those interested in implementations are referred to the web services of the European Bioinformatics Institute[1] and related sites.

23.2.2. Compactness

A variety of ingenious techniques have been invented for locating structural domains in 3D structures. These include inspection of distance maps, clustering, neighbourhood correlation, plane cutting, interface area minimization, specific volume minimization, searching for mechanical hinge points, maximization of compactness and maximization of buried surface area (Rossmann & Liljas, 1974; Rashin, 1976; Crippen, 1978; Nemethy & Scheraga, 1979; Rose, 1979; Schulz & Schirmer, 1979; Go, 1981; Lesk & Rose, 1981; Sander, 1981; Wodak & Janin, 1981; Zehfus & Rose, 1986; Kikuchi *et al.*, 1988; Moult & Unger, 1991; Holm & Sander, 1994; Zehfus, 1994; Islam *et al.*, 1995; Siddiqui & Barton, 1995; Swindells, 1995; Holm & Sander, 1996; Sowdhamini *et al.*, 1996; Zehfus, 1997; Holm & Sander, 1998; Jones *et al.*, 1998; Wernisch *et al.*, 1999).

Common to most approaches are the assumptions that folding units are compact and that the interactions between them are weak. These notions can be made quantitative, for example, by counting interatomic contacts and by locating domain borders by identifying groups of residues such that the number of contacts between groups is minimized. The hierarchic organization of putative folding units can be inferred starting from the complete structure and recursively cutting it (*in silico*) into smaller and smaller substructures. Alternatively, one may start from the residue or secondary-structure-element level and successively associate the most strongly interacting groups. The procedure involves two optimization problems.

The first optimization problem is algorithmic and concerns finding the optimal subdivisions. This problem is complicated by the possibility of the chain passing several times between domains (discontinuous domains). Without the constraint of sequential continuity, there is a combinatorial number of possibilities for dividing a set of residues into subsets (Zehfus, 1994). This hurdle has been overcome by fast heuristics (Holm & Sander, 1994*b*; Zehfus, 1997; Wernisch *et al.*, 1999).

The second optimization problem concerns formulating physical criteria that distinguish between autonomous and nonautonomous folding units, *i.e.*, defining termination criteria for recursive algorithms. Since compactness-related criteria do not have a clear bimodal distribution, domain-assignment algorithms (Holm & Sander, 1994*b*; Islam *et al.*, 1995; Siddiqui & Barton, 1995; Swindells, 1995; Sowdhamini *et al.*, 1996; Wernisch *et al.*, 1999) use cutoff parameters that have been fine-tuned against an external reference set of domain definitions.

23.2.3. Recurrence

Most fold classifications use a hierarchical model where evolutionary families are a subcategory of fold type and it is natural to assume that domain boundaries should be conserved in evolution. Consistency concerns lead to a reformulation of the goals of the domain-assignment problem, away from (imprecise) physical models of stable folding units and towards recognizing such units phenomenologically in the database of known structures through recurrence. The concept of recurrence has long been the cornerstone of domain assignments by experts based on visual inspection (Richardson, 1981). Recurrence means recognizing architectural units in one protein that have already been defined (named) in another.

The practical importance of domain identification is illustrated by the discoveries made by a systematic structure comparison of recurrent domains between histidine triad (HIT) proteins and galactose-6-phosphate uridylyltransferase [homodimer and internally duplicated common catalytic core, respectively (Holm & Sander, 1997)], and between beta-glucosyltransferase and glycogen phosphorylase [bare and heavily decorated common catalytic core, respectively (Holm & Sander, 1995; Artymiuk *et al.*, 1995)], even though the contours of the molecules look quite different.

Let us restate the goal of domain identification as an economic description of all known protein structures in terms of a small set of large substructures. This is an intuitive goal and conceptually related to the principle of minimal encoding in information theory. The key ingredients of the optimization problem are the gain associated with reusing a substructure and the cost associated with using many small substructures to describe a protein. An analogy in writing is that copying blocks of text is cheap, but for coherence some thought and effort is necessary for bridging the blocks.

---

[1] EMBL–EBI (1995): http://www.ebi.ac.uk/; DALI domain dictionary (1999): http://www.ebi.ac.uk/dali/domain/; 3Dee – database of protein domain definitions (1997): http://barton.ebi.ac.uk/servers/3Dee.html.

Copyright © 2012 International Union of Crystallography

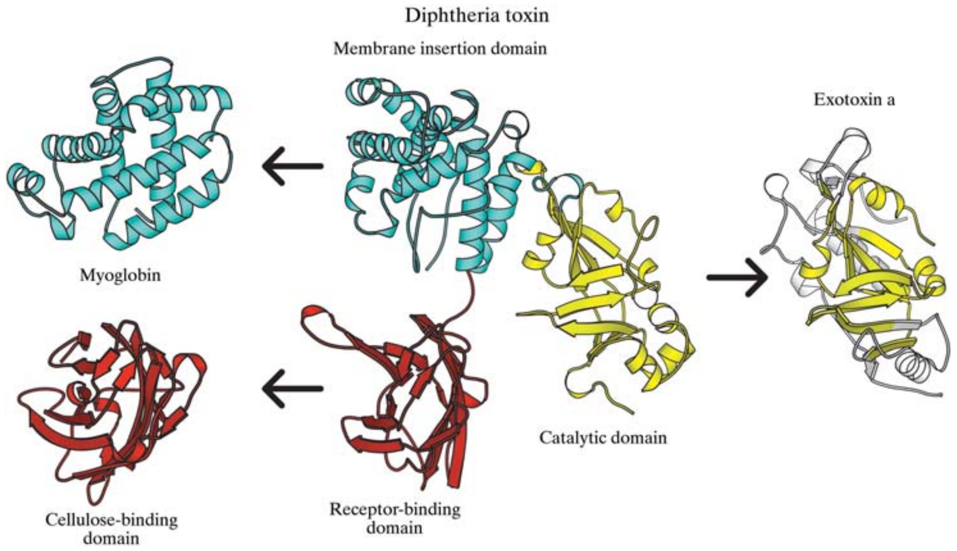


Figure 23.2.1.1
The structure of diphtheria toxin (Bennett & Eisenberg, 1994) beautifully illustrates domains as structural, functional and evolutionary units. Structurally, note the compact globular shape of each domain and the flexible linkers between them. Functionally, note how each domain carries out a different stage of infection by the bacterium: receptor binding, membrane penetration and ADP-ribosylation of the target protein. Evolutionarily, note the occurrence of domains homologous to the catalytic domain of diphtheria toxin in exo-, entero- and pertussis toxins, and in poly-ADP-ribose polymerase (Holm & Sander, 1999). Arrows point to recurrent substructures in structural neighbours (Lionetti *et al.*, 1991; Li *et al.*, 1996; Tormo *et al.*, 1996) of each domain of diphtheria toxin. Drawn using *MOLSCRIPT* version 2 (Kraulis, 1991).

With a suitably defined cost function, recurrence can be used to select an optimal set of substructures from the hierarchic folding or unfolding trees generated using compactness criteria. Thus, the unsatisfactorily solved problem of defining termination criteria for compactness algorithms can be turned into an optimization problem that does not rely on any external reference and leads to an internally consistent set of domain definitions.

The key difficulty is in quantifying the notion of economy so that it leads to a selection of substructures of 'appropriate' size, *i.e.*, globular folds and not, for example, supersecondary-structure motifs. One solution, which is physical nonsense but has the desired qualitative behaviour, is a heuristic objective function used in the DALI domain dictionary (Holm & Sander, 1998). Recurrence is quantified in terms of the statistical significance of structural similarity for many pairs of substructures. The statistical significance is highest for structural similarities that involve large units and that completely cover a substructure unit. Exploiting these effects, a sum-of-pairs objective function is defined that favours recurrences of large substructures with distinct topological arrangements and packing of secondary-structure elements, and disfavours small substructures consisting of one or two secondary-structure elements despite their higher frequency of recurrence. Though other formulations of the optimization problem are possible, this empirically chosen objective function combined with a heuristic algorithm for optimization yields a useful set of substructures (domains).

23.2.4. Conclusion

While we do not foresee that automatically delineated domains will be accepted as the gold standard of the trade, modern methods, based on a combination of recurrence and compactness criteria, yield domain definitions that are consistent within protein families and often coincide with biologically functional units, recover the well known folding topologies with many members, produce clusters with good coverage of common secondary-structure elements, and provide a useful basis for large-scale structure analysis and classification.

References

Artymiuk, P. J., Rice, D. W., Poirrette, A. R. & Willett, P. (1995). *Beta-glucosyltransferase and phosphorylase reveal their common theme*. *Nat. Struct. Biol.* **2**, 117–120.

Bennett, M. J. & Eisenberg, D. (1994). *Refined structure of monomeric diphtheria toxin at 2.3 Å resolution*. *Protein Sci.* **3**, 1464–1475.

Bork, P. (1992). *Mobile modules and motifs*. *Curr. Opin. Struct. Biol.* **2**, 413–421.

Crippen, G. (1978). *The tree structural organization of proteins*. *J. Mol. Biol.* **126**, 315–332.

Go, M. (1981). *Correlation of DNA exonic regions with protein structural units in hemoglobin*. *Nature (London)*, **291**, 90–92.

Holm, L. & Sander, C. (1994). *Parser for protein folding units*. *Proteins*, **19**, 256–268.

Holm, L. & Sander, C. (1995). *Evolutionary link between glycogen phosphorylase and a DNA modifying enzyme*. *EMBO J.* **14**, 1287–1293.

Holm, L. & Sander, C. (1996). *Mapping the protein universe*. *Science*, **273**, 595–602.

Holm, L. & Sander, C. (1997). *Enzyme HIT*. *Trends Biochem. Sci.* **22**, 116–117.

Holm, L. & Sander, C. (1998). *Dictionary of recurrent domains in protein structures*. *Proteins*, **33**, 88–96.

Holm, L. & Sander, C. (1999). *Protein folds and families: sequence and structure alignments*. *Nucleic Acids Res.* **27**, 244–247.

Islam, S. A., Luo, J. & Sternberg, M. J. (1995). *Identification and analysis of domains in proteins*. *Protein Eng.* **8**, 513–525.

Jones, S., Stewart, M., Michie, A. D., Swindells, M. B., Orengo, C. A. & Thornton, J. M. (1998). *Domain assignment for protein structures using a consensus approach: characterisation and analysis*. *Protein Sci.* **7**, 233–242.

Kikuchi, T., Nemethy, G. & Scheraga, H. A. (1988). *Prediction of the location of structural domains in globular proteins*. *J. Protein Chem.* **88**, 427–471.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures*. *J. Appl. Cryst.* **24**, 946–950.

Lesk, A. M. & Rose, G. D. (1981). *Folding units in globular proteins*. *Proc. Natl Acad. Sci. USA*, **78**, 4304–4308.

Li, M., Dyda, F., Benhar, I., Pastan, I. & Davies, D. R. (1996). *Crystal structure of the catalytic domain of Pseudomonas exotoxin A complexed with a nicotinamide adenine dinucleotide analog: implications for the activation process and for ADP ribosylation*. *Proc. Natl Acad. Sci. USA*, **93**, 6902–6906.

Lionetti, C., Guanziroli, M. G., Frigerio, F., Ascenzi, P. & Bolognesi, M. (1991). *X-ray crystal structure of the ferric sperm whale myoglobin: imidazole complex at 2.0 Å resolution*. *J. Mol. Biol.* **217**, 409–412.

Moult, J. & Unger, R. (1991). *An analysis of protein folding pathways*. *Biochemistry*, **30**, 3816–3824.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of the protein database for the investigation of sequences and structures*. *J. Mol. Biol.* **247**, 536–540.

Nemethy, G. & Scheraga, H. A. (1979). *A possible folding pathway of bovine pancreatic Rnase*. *Proc. Natl Acad. Sci. USA*, **76**, 6050–6054.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *CATH – a hierarchic classification of protein domain structures*. *Structure*, **5**, 1093–1108.

Rashin, A. A. (1976). *Location of domains in globular proteins*. *Nature (London)*, **291**, 85–87.

Richardson, J. S. (1981). *The anatomy and taxonomy of protein structure*. *Adv. Protein Chem.* **34**, 167–339.

Rose, G. D. (1979). *Hierarchic organization of domains in globular proteins*. *J. Mol. Biol.* **134**, 447–470.

Rossmann, M. & Liljas, A. (1974). *Recognition of structural domains in globular proteins*. *J. Mol. Biol.* **85**, 177–181.

Sander, C. (1981). *Physical criteria for folding units of globular proteins*. In *Structural Aspects of Recognition and Assembly in Biological Macromolecules*, Vol. I. *Proteins and Protein Complexes, Fibrous Proteins*, edited by M. Balaban, pp. 183–195. Jerusalem: Alpha Press.

Schulz, G. E. & Schirmer, H. (1979). *Principles of Protein Structure*, ch. 5. New York: Springer Verlag.

Siddiqui, A. S. & Barton, G. J. (1995). *Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions*. *Protein Sci.* **4**, 872–884.

Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *A database of globular protein structural domains: clustering of representative family members into similar folds*. *Structure Fold. Des.* **1**, 209–220.

Swindells, M. B. (1995). *A procedure for detecting structural domains in proteins*. *Protein Sci.* **4**, 103–112.

Tormo, J., Lamed, R., Chirino, A. J., Morag, E., Bayer, E. A., Shoham, Y. & Steitz, T. A. (1996). *Crystal structure of a bacterial family-III cellulose-binding domain: a general mechanism for attachment to cellulose*. *EMBO J.* **15**, 5739–5751.

Wernisch, L., Hunting, M. & Wodak, J. (1999). *Identification of structural domains in proteins by a graph heuristic*. *Proteins*, **35**, 338–352.

Wetlaufer, D. B. (1973). *Nucleation, rapid folding, and globular intrachain regions in proteins*. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.

Wodak, J. & Janin, J. (1981). *Location of structural domains in proteins*. *Biochemistry*, **20**, 6544–6552.

Zehfus, M. H. (1994). *Binary discontinuous compact protein domains*. *Protein Eng.* **7**, 335–340.

Zehfus, M. H. (1997). *Identification of compact, hydrophobically stabilized domains and modules containing multiple peptide chains*. *Protein Sci.* **6**, 1210–1219.

Zehfus, M. H. & Rose, G. D. (1986). *Compact units in proteins*. *Biochemistry*, **25**, 5759–5765.

International Tables for Crystallography (2012). Vol. F, Chapter 23.3, pp. 755–765.

Chapter 23.3. Protein–ligand interactions

A. E. Hodel and F. A. Quiocho

23.3.1. Introduction

There are currently over a thousand unique protein-bound ligands described in the Protein Data Bank, illustrating the enormous variety of small molecules with which proteins interact. These ligands serve as cofactors in protein-mediated reactions, substrates in these reactions, and elements that maintain or alter protein structure or macromolecular assembly. The specific binding of small molecules to proteins is a primary means by which living systems interact and exchange information with their environment. The atomic details of protein–ligand interactions are often quite similar to the intramolecular interactions observed within a protein molecule. Examples of all the various non-covalent interactions described in Parts 20 and 22, such as hydrogen bonds, van der Waals forces and other electrostatic phenomena, are observed between proteins and their small-molecule ligands.

Through the diverse interactions observed between proteins and their ligands, a few fundamental patterns of recognition emerge. In general, ligand binding requires that the protein partially or fully sequesters the ligand from the solvent. This demands that the energy of interaction between the protein and the ligand must be strong enough to overcome the interactions between both species and the solvent as well as the translational and rotational entropy which is lost upon fixing the orientation of the ligand relative to the protein. The protein achieves this level of interaction by presenting a binding site that is complementary to the ligand both in shape and electrostatic functionality. Beyond this generalization, each ligand has its own unique and complicated story. Rather than attempt to summarize the enormous subject of protein–ligand interactions in a comprehensive manner, we will instead illustrate several of the unique interactions observed between proteins and other molecules.

23.3.2. Protein–carbohydrate interactions

The interaction between proteins and carbohydrates provides a prototypic example of how proteins specifically recognize small organic ligands. Protein-mediated recognition of carbohydrates is crucial in a diverse array of processes, including the transport, biosynthesis and storage of carbohydrates as an energy source, signal transduction through carbohydrate messengers, and cell–cell recognition and adhesion (Rademacher *et al.*, 1988). Many aspects of protein–carbohydrate recognition are observed in the interactions of proteins with other small-molecule ligands. For example, carbohydrate recognition is largely conferred through a combination of hydrogen bonding and van der Waals interactions with the protein (Quiocho, 1986). These interactions are generally presented in a binding site that is highly complementary to the target ligand in shape and functionality. Carbohydrate-binding proteins also occasionally employ ordered water molecules and bound metal ions to facilitate ligand recognition (Quiocho *et al.*, 1989; Vyas, 1991; Weis & Drickamer, 1996). There are also many examples of 'induced fit' recognition of carbohydrates, where the binding of the ligand induces a conformational change in the protein. This phenomenon of ligand-induced conformational change was observed in one of the first ligand–protein interactions visualized by X-ray crystallography, namely the binding of a carbohydrate substrate to lysozyme (Blake *et al.*, 1967). Since these early studies, a wide variety of carbohydrate–protein structures have been determined, and a number of general themes have emerged from this continually active field (Vyas, 1991). These themes are further applicable in the general study of protein–ligand recognition.

Two general classes of carbohydrate-binding modes have been observed in their complexes with proteins (Quiocho, 1986; Vyas, 1991). The first group of proteins completely sequester the carbohydrate from the surrounding solvent. These proteins, including the periplasmic proteins and the catalytic site of glycogen phosphorylase, tend to have a high affinity for their ligand ($K_d \simeq 10^{-6}$–10^{-7}). The high affinity of these proteins can be attributed to the entropic gain of desolvating the protein and carbohydrate surfaces as well as the exceptionally high degree of functional complementarity in the ligand-binding site. The periplasmic proteins nearly saturate the hydrogen-bonding potential of their carbohydrate ligands (Quiocho, 1986). The second group of proteins bind their ligands in shallow clefts in the solvent-exposed protein surface. This mode of binding, observed in the lectins, lysozyme and the storage site of glycogen phosphorylase, generally has a lower binding constant than the first class of binding proteins ($K_d \simeq 10^{-3}-10^{-6}$). Some members of this second class of proteins, such as the lectins, are able to increase the affinity for their ligands dramatically by clustering a number of low-affinity sites through the oligmerization of the polypeptides (Weis & Drickamer, 1996). The specific arrangement of lectin oligomers allows them to discriminate between large cell-surface arrays of polysaccharides with high affinity and selectivity.

23.3.2.1. Carbohydrate recognition at the atomic level

An early review of protein–carbohydrate interactions revealed several atomic level interactions that continue to appear ubiquitously in the structures of protein–carbohydrate complexes (Quiocho, 1986). A generic cyclic sugar in mono- or oligosaccharides appears to be recognized as a disc that displays two flat non-polar surfaces surrounded by a ring of polar hydroxyls. Proteins recognize these features by hydrogen bonding to the ring of polar hydroxyls while stacking flat aromatic side chains against the non-polar disc faces.

Cooperative hydrogen bonding, where the hydroxyl group of a carbohydrate participates as both a donor and acceptor of hydrogen bonds, is often observed in the direct interactions between proteins and carbohydrate ligands (Fig. 23.3.2.1). The sp^3-hybridized oxygen atom of a carbohydrate hydroxyl may act as both an acceptor of two hydrogen bonds through the two lone pairs of electrons as well as a donor of a single hydrogen bond. Cooperative hydrogen bonding generally follows a simple pattern in which the carbohydrate hydroxyl accepts a hydrogen bond from a protein amide group while simultaneously donating a hydrogen bond to a protein carbonyl oxygen. Hydrogen bonding to protein hydroxyl groups is observed only infrequently.

Copyright © 2012 International Union of Crystallography

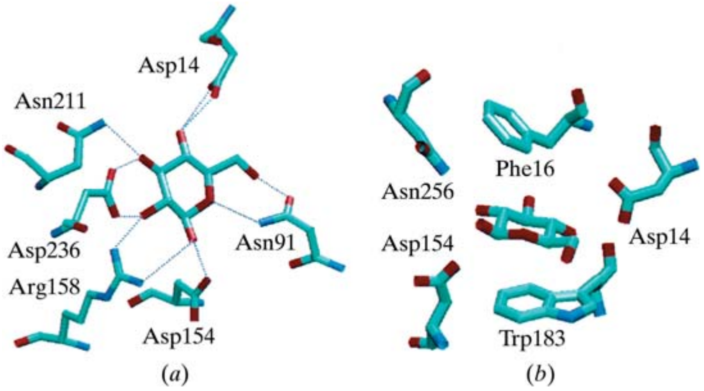


Figure 23.3.2.1
The atomic recognition of carbohydrates by a protein. The environment of glucose bound to the galactose/glucose binding protein is shown (Vyas *et al.*, 1988). (*a*) The ring of hydrogen bonds around the polar edge of the sugar molecule. Note the 'bidentate' hydrogen bonds with Asp236, Arg158 and Asn91. As the hydroxyl groups 1, 2 and 3 simultaneously donate and accept hydrogen bonds from protein side chains, they are involved in 'cooperative' hydrogen bonds. (*b*) Two aromatic protein side chains, Phe16 and Trp183, 'stack' on the non-polar faces of the carbohydrate.

This pattern is thought to be a result, in part, of the entropic cost of fixing a freely rotating protein hydroxyl group while simultaneously fixing the ligand hydroxyl group. Amides and carbonyls are usually fixed in a planar geometry and thus do not require as much energy to compensate for their loss of entropy in ligand binding.

The vicinal hydroxyl groups of carbohydrates provide an ideal geometry for the formation of 'bidentate' hydrogen bonds, where the pair of hydroxyls interacts with two functional groups of a single amino-acid side chain or the main-chain amide groups of two consecutive residues (Fig. 23.3.2.1). These interactions occur when the adjacent carbohydrate hydroxyls are either both equatorial, or one is equatorial and the other axial. The inter atomic distance for the carbohydrate hydroxyl oxygens is ~2.8 Å in these cases, allowing for a bidentate interaction with the planar side chains of aspartate, asparagine, glutamate, glutamine and arginine. Bidentate hydrogen bonds have not been observed for consecutive axial hydroxyls where the oxygen–oxygen distance increases to ~3.7 Å.

Carbohydrates often present a disc-like face of non-polar aliphatic hydrogen atoms which proteins recognize through the use of aromatic side chains. The protein aromatic groups are 'stacked' on the flat face of the carbohydrate, thus generating both specificity and binding energy through van der Waals interactions. Tryptophan, the aromatic amino acid with the largest surface area and highest electronegativity, is the most common side chain employed in van der Waals 'stacking' with carbohydrates. The infrequent use of aliphatic groups in the binding of the non-polar carbohydrate faces suggests that the aromatic moieties are employed in a specific manner. The electron-rich electron clouds of the aromatic side chains may provide a strong electrostatic interaction with the aliphatic carbohydrate protons that could not be satisfied by protein aliphatic groups. The anionic character of aromatic side chains is observed in a number of protein–intramolecular (Chapter 22.3) and protein–ligand interactions (see below).

23.3.3. Metals

Metal ions provide a number of important functions in their diverse and ubiquitous interactions with proteins. The most common function for a protein-bound metal ion is the stabilization and orientation of the protein tertiary structure through coordination to specific protein functional groups. In addition to this structural role, metal ions are also often directly involved in enzyme catalysis and protein function. Examples of these functions include redox reactions, the activation of chemical bonds and the binding of specific ligands. Myoglobin, the first protein structure determined by X-ray crystallography, specifically binds molecular oxygen through an iron ion of a haem cofactor. Myoglobin provides a prototypic example of a protein and a metal ion providing a unique and specific functionality through their combination.

23.3.3.1. Metals important in protein function and structure

A number of metals are relatively abundant and available in living systems (Table 23.3.3.1) (Glusker, 1991). The most common ions include sodium, potassium, magnesium and calcium. Along with these ions, a large variety of trace metals are also found coordinated to proteins. The structures of protein complexes with some of these trace ions, including iron, zinc and copper, have been studied extensively for some time (Glusker, 1991). More recently, the structures of protein complexes with more unusual ions, such as nickel, vanadium and tungsten, have been determined (Volbeda *et al.*, 1996).

Specificity in the interactions between proteins and metal ions is conferred through each ion's preference for the coordinating atoms and the geometry of the binding site. All four of the more common metals, *i.e.* sodium, potassium, magnesium and calcium, are classified as 'hard' metals, referring to the polarizability of the

Table 23.3.3.1
Metal ions associated with proteins

Metal ion	Concentration in blood plasma (m*M*)	Common cofactors	Hard/soft classification	Common coordination number and geometry	Preferred ligand atom
Na^+	138		Hard	6	O
K^+	4		Hard	8	O
Ca^{2+}	3		Hard	8	O
Mg^{2+}	1		Hard	6	O
Fe	0.02	Haem	Intermediate	6 (octahedral)	N
Zn^{2+}	0.02		Intermediate	4 (tetrahedral), 6 (octahedral)	N, S
Cu^{2+}	0.015		Soft	4 (tetrahedral)	S
Co^{2+}	0.002		Hard	6 (octahedral)	O
Mn^{2+}	0		Hard	6 (octahedral)	O
Ni^{2+}	0		Intermediate	6 (octahedral)	N
Mo	0	Pterin	Intermediate	6 (octahedral)	S
W	0	Pterin	Intermediate	6 (octahedral)	S
V	0			5 (trigonal bipyramidal)	

electron cloud of the ion. The nucleus of a hard metal has a relatively tight hold on the surrounding electrons. These ions lack easily excitable unshared electrons and have a low polarizability. The interactions between these metals and their ligands tend to have the character of ionic interactions rather than the more covalent nature preferred by the 'soft' metals. In general, the hard metals prefer to coordinate with hard acids, such as the oxygen atoms of hydroxyls, carbonyls and carboxyls.

The soft metals have a high polarizability, large ionic radius and several unshared valence electrons. They generally prefer to coordinate with soft acids, such as the thiol and thiol ether groups of cysteine and methionine. The loosely held valence electrons of soft metals tend to favour partially covalent π-bonding with their coordinated ligands. These outer-shell electrons can be donated to the empty outer orbitals of the ligand atom. The partially covalent nature of these bonds yields more stable complexes than the ionic complexes of the hard metals. This partial covalent bond also polarizes the ligand coordinated to the metal and can thus activate adjacent atoms to nucleophilic attack.

A large number of the transition metals, including zinc and iron, form ions that have intermediate polarizability with regard to hard and soft metals. These ions mainly prefer nitrogen ligands like the imidazole side chain of histidine or the central nitrogens of the haem cofactor.

The geometry of the metal-binding site in a protein depends on a combination of the radial size of the ion as well as the polarizability of the metal. The number of coordinating ligands around the metal is primarily correlated with the relative size of the ion, where as many anions as possible are packed around the cationic metal without leaving any cavities (Orgel, 1966). This leads to a relatively simple correlation between the ratio of the radii of the cation and the anion ($r_{\rm cation}/r_{\rm anion}$) with the coordination number. Beyond this simple geometric constraint, the coordination number is also influenced by the repulsion between the closely packed anion ligands. This repulsion can be tempered by the distortions in the cation's electron cloud, leading to a dependency between the coordination number and the polarizability of the metal ion. Table 23.3.3.1 gives the most common coordination numbers and geometries for the listed metal ions. For a more comprehensive description of possible coordination geometries, see Glusker (1991).

A short example of the diversity of metal functions in protein complexes is found in a comparison between the calcium-binding proteins calmodulin and staphyloccocal nuclease. Calmodulin functions in signal transduction by binding to a wide variety of proteins in a calcium-dependent manner. In the absence of calcium, calmodulin adopts a conformation where two loosely folded domains are connected by a flexible α-helix analogous to two balls tied together by a string. In the presence of Ca^{2+}, each of the two domains of calmodulin binds to a single metal ion. The binding of Ca^{2+} to the two calmodulin domains induces a large conformational change in the protein, which confers a high affinity for peptide ligands. Crystallographic studies show that the two calcium-bound domains form a clamp that closes on the target peptide ligand (Meador *et al.*, 1995). Thus, in this case, the metal ion plays an indirect role as a structural element in the protein function.

In the case of staphylococcal nuclease, calcium binding appears to play a more direct role in the catalytic function of the protein. A Ca^{2+} ion binds at the active site and coordinates with protein side chains, water molecules and the substrate phosphate group. The addition of calcium affects the nuclease reaction both in the binding of the substrate and directly in the catalytic step. Although calcium increases the K_m of the nucleic acid substrate, this effect can be reproduced with a large number of other metal ions (Tucker *et al.*, 1979). The effect on catalysis, however, is specific to Ca^{2+} ions. In a proposed mechanism, Ca^{2+} directly contributes to catalysis by activating a water-derived hydroxide ion for nucleophilic attack on the phosphorus atom of the nucleic acid backbone (Cotton *et al.*, 1979).

23.3.4. Protein–nucleic acid interactions

23.3.4.1. The DNA double helix

DNA provides one of the more compelling protein 'ligands' for biophysical study, as the sequence-specific binding of proteins to the DNA double helix mediates the interaction between the environment surrounding the living cell and the information 'programmed' into the cell within its genome. A classic example of such a process is the response of the bacteria *Escherichia coli* to the nutrients in the surrounding media through the regulation of gene expression. A simple case of this interaction is found in the biosynthesis of the amino acid tryptophan. The transcription of the genes necessary for the synthesis of tryptophan is suppressed when tryptophan is present in the environment. This process is mediated by the tryptophan-dependent sequence-specific binding of the trp repressor protein to the *trp* operon within the genes encoding the metabolic enzymes (Joachimiak *et al.*, 1983). In the absence of tryptophan, the affinity of the aporepressor for the *trp* operon is dramatically reduced. Thus, when tryptophan is not available in the environment, transcription of the biosynthetic genes proceeds. In mammalian cells, the analogous process is observed in the activation of gene expression through hormones, cytokines and other stimuli.

Although DNA has often been considered to be a long, nearly featureless cylindrical double helix, proteins have evolved with exquisite specificity for their cognate DNA sequences. This apparent contradiction can be reconciled with the acknowledgement of two recently appreciated properties of DNA (Harrington & Winicov, 1994). First, the local structure of DNA is actually highly variable and dependent on the specific sequence of the base pairs in the helical ladder. Second, the DNA double helix is a relatively soft structure that is easily deformed into concerted bends, kinks and other distortions. DNA-binding proteins thus recognize their cognate sequences both by utilizing the unique local structure of the double helix and by inducing distortions into the helix which facilitate recognition.

The most intuitive features of the double helix that are important in sequence-specific recognition are the unique surfaces presented by the bases in the helix grooves. DNA is primarily found in a B-form helix that presents a wide, accessible major groove and a deep, narrow minor groove. An analysis of the arrangement of hydrogen-bonding functional groups presented by DNA bases (Fig. 23.3.4.1) suggests that the sequence-specific recognition of the DNA helix is best facilitated through the major groove, where each of the four possible base-pair combinations present unique hydrogen-bonding patterns (Steitz, 1990). The majority of sequence-specific DNA-binding proteins of known structure appear to utilize this direct readout of the major groove by inserting a portion of an α-helix, a two-stranded β-hairpin, or even a peptide coil which presents complementary hydrogen-bonding arrangements with the DNA bases (Pabo & Sauer, 1992; Steitz, 1990). The narrow surface of the minor groove presents some characteristic hydrogen-bonding patterns: however, the absolute identity of each base pair is

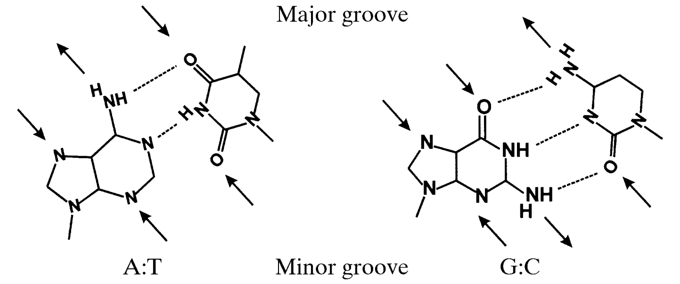


Figure 23.3.4.1
A schematic diagram of the base pairs of DNA showing the hydrogen-bonding groups which may be used in the sequence-specific recognition of DNA. The major groove is at the top of the figure and the minor groove at the bottom. Arrows point towards hydrogen-bond acceptors and away from donors.

ambiguously represented in these patterns (Fig. 23.3.4.1). The similar position of hydrogen-bonding groups in the minor groove would make it hard to distinguish AT base pairs from TA base pairs and GC base pairs from CG base pairs. Although there are proteins that recognize DNA through the minor groove, such as the TATA-box binding protein, the recognition of their target is completed through dramatic distortion of the DNA helix through intercalation (see below).

α-Helices are the most frequently observed structural motif for recognition in the major groove of DNA (Pabo & Sauer, 1992). The overall shape and dimensions of the α-helix are geometrically suited for binding in the major groove of a B-DNA helix (Fig. 23.3.4.2). The exact orientations of helices in various protein–DNA complexes are quite variable. Most helices bind in the major groove at an angle of approximately 30 (15)° from the plane normal to the DNA helical axis (Fig. 23.3.4.3). However, the numerous variants to this rule would include the trp repressor/operator complex, where only the N-terminal end of the 'recognition' helix is inserted into the major groove (Otwinowski *et al.*, 1988*a*,*b*). Interactions observed between these inserted elements and the DNA bases include the common direct hydrogen bond between the protein side chain and base, the less common hydrogen bond between the protein backbone and base, indirect but specific hydrogen bonding through water molecules, and hydrophobic interactions.

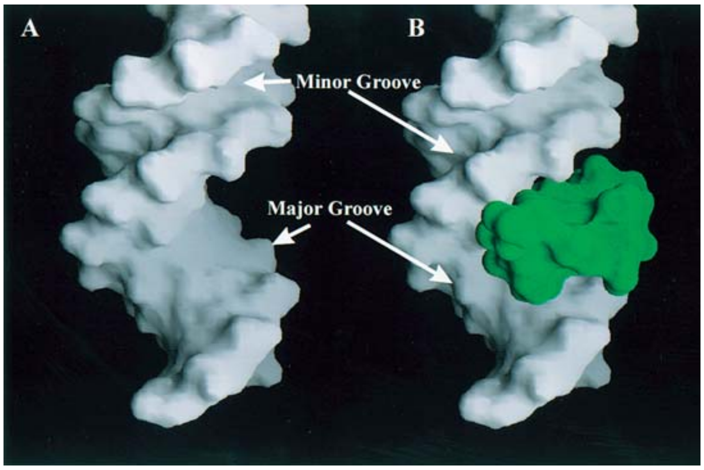


Figure 23.3.4.2
(*a*) A space-filling model of B-DNA showing the relative accessibility of the major and minor grooves. (*b*) A helix of the 434 repressor bound in the major groove of the helix, illustrating how the dimensions of a protein α-helix are compatible for reading the major groove of B-DNA (Shimon & Harrison, 1993).

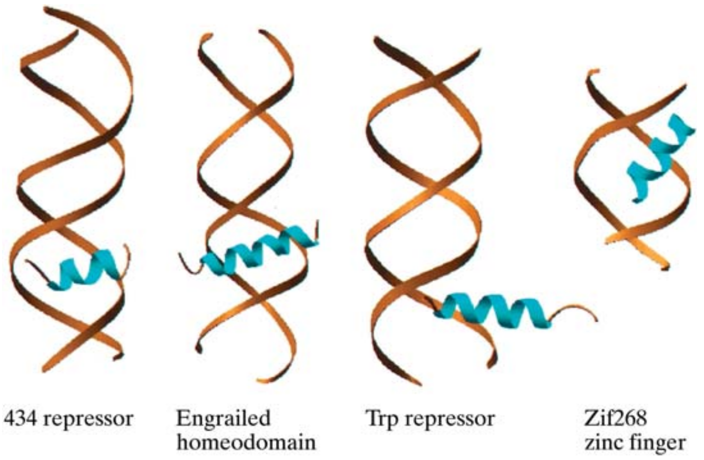


Figure 23.3.4.3
A comparison of the orientations of α-helices bound in the major groove, taking examples from four DNA-binding proteins: the 434 repressor (Shimon & Harrison, 1993), the engrailed homeodomain (Kissinger *et al.*, 1990), the trp repressor (Otwinowski *et al.*, 1988*a*,*b*) and the Zif268 zinc finger (Pavletich & Pabo, 1991). The DNA backbone is shown as a brown ribbon, whereas the protein helix is shown as a blue ribbon.

There appears to be no simple correlation between the primary sequence of the peptide segments which make specific base contacts and the DNA sequence that those segments recognize (Pabo & Sauer, 1992; Steitz, 1990). Examples of every polar protein side chain participating in specific hydrogen bonds with DNA bases have been observed, but each amino acid does not show any preference for any one particular base. What is observed is that conserved residues within families of DNA-binding proteins tend to make conserved base-specific interactions in DNA–protein complexes. Strikingly, this subset of interactions which are conserved within protein families include cooperative hydrogen bonding reminiscent of the pairs of hydrogen bonds often observed in carbohydrate–protein complexes. These interactions, which include the pairing of arginine with guanine and glutamine or asparagine with adenine, were predicted early on by Seeman *et al.* (1976).

Although the elements of protein structure in direct contact with the DNA bases play a prominent role in sequence specificity, these elements are not sufficient to impart the specificity of the DNA-binding protein. This statement is supported by the variety of orientations in which the 'recognition' helices bind to the major groove. The structural context of the recognition elements and the overall docking of the protein to the DNA helix play as important a role in specificity as the direct base interactions.

The contacts between the protein and the ribose–phosphate backbone of the DNA appear to be one of the more important aspects of the 'indirect readout' of the DNA sequence (Pabo & Sauer, 1992). On average, more than half of the interactions between protein and DNA in complex structures involve the backbone of the DNA helix. Thus, the sheer number of interactions suggests that these contacts serve an important function in recognition. Although several of the protein–DNA-backbone contacts observed involve salt bridges between the phosphates and basic protein side chains, these interactions are not as highly represented as one might expect. This could be a result of the high degree of flexibility inherent in the long side chains of arginine and lysine. Instead, examples of every basic and neutral residue and occasionally even acidic residue with some hydrogen-bonding potential interacting with the phosphate backbone have been observed. These contacts may contribute to specificity through two mechanisms. First, they can establish the exact orientation of the base-specific contacts relative to the 'rungs' in the phosphate backbone. Second, they may read the base sequence indirectly through sequence-specific backbone distortions or flexibility. There are numerous examples of DNA–protein complexes with highly distorted DNA helices. There is also evidence that certain DNA sequences inherently confer bends within the B-form helix. Thus, it is conceivable that protein interactions with the DNA backbone may confer specificity by selecting for a specific distorted conformation of the helix.

The most dramatic distortion of the DNA helix has been observed in DNA–protein complexes where the protein induces a kink or bend through the intercalation of the DNA helix at the minor groove (Werner *et al.*, 1996*a*,*b*). Intercalation involves the insertion of a hydrophobic protein side chain into the helix, disrupting the stacking of two adjacent base pairs, and, in some cases, the side chain itself then stacks with one of the base pairs. Examples of this mode of binding include the complexes of the TATA-box binding protein (TBP), the PurR repressor and the human oncogene *ETS1* with their cognate DNA partners (Werner *et al.*, 1996*a*,*b*). The *ETS1*–DNA complex provides the only current example of complete intercalation of the DNA extending from the minor groove to the major groove. A tryptophan side chain extends into the helix from the minor groove and stacks with one of the displaced base pairs. The remaining base pair contacts the ring system of the tryptophan edge in forming a pseudo-hydrogen bond between the indole hydrogens and the π-rings of the DNA bases. In *ETS1*, the deformation of the DNA helix resulting from protein intercalation results in the kinking of the helical axis from 45° to about 60°.

Examples of protein intercalation of the DNA helix from the major groove are found in proteins, such as the methyltransferases, that perform chemistry on the bases of the DNA. To perform their enzymatic function, these proteins must extract the target base from the DNA helix and 'flip' the base out into the enzyme active site (Cheng, 1995). The resulting void in the DNA is then filled by protein side chains that partially satisfy the hydrogen-bonding and van der Waals interactions that were broken when the target base was flipped. Although there are only a few known structures of DNA–protein complexes with extrahelical bases, base flipping is thought to be a relatively common feature of DNA-modifying enzymes.

23.3.4.2. *Single-stranded sequence-nonspecific DNA–protein interactions*

There have been a few reports of single-stranded DNA–protein complex structures, all of which involve the sequence-nonspecific recognition of DNA. In the binding of a tetranucleotide to the exonuclease active site of the DNA polymerase I Klenow fragment (Freemont *et al.*, 1988), extensive hydrogen-bonding interactions between the sugar–phosphate backbone and the protein are observed. This provides the most intuitive mechanism for sequence-nonspecific nucleic acid binding, where the protein simply recognizes the phosphate backbone of a single-stranded coil. The protein also appears to form a few hydrophobic interactions with the DNA bases; however, these interactions, which include the partial intercalation between two bases, are thought to be nonspecific.

The structure of replication protein A complexed with single-stranded DNA does not exhibit the intuitive nonspecific mechanism of recognition found in the Klenow fragment (Bochkarev *et al.*, 1997). In this structure, the DNA is extended with its bases splayed out over the surface of the protein. The bases form several pairwise stacking interactions that are interrupted by intercalating protein side chains. Contrary to the sequence-nonspecific nature of recognition, numerous hydrogen bonds are found between the protein and the bases of the DNA strand. These base-dependent contacts require that the protein–DNA interactions must be flexible and plastic in order to accommodate different base sequences.

23.3.4.3. *RNA*

Although RNA and DNA are chemically similar, RNA presents a much greater variety of shapes and surfaces compared to the relatively simple B-form helix of DNA. Generally single-stranded, RNA often forms secondary structuresdriven by the base pairing of complementary stretches of sequence within the same strand. The formation of base-paired regions can result in stem loops, bulges and helices which can further assemble into more complicated tertiary structures, such as that observed for transfer RNAs. Protein-mediated recognition of RNA often depends as much on the three-dimensional structure presented by these secondary structures as on the specific identity of the base sequence.

Very little information is currently available on the structural details of protein–RNA interactions (Nagai, 1996). Only a handful of protein–RNA complex structures have been determined. These fall into three basic categories, depending on the secondary structure of the RNA: four tRNA–protein complexes, two stem-loop–protein complexes and a capped single-stranded RNA–protein complex.

23.3.4.4. *Transfer RNA*

In the four known structures of tRNA bound to their aminoacyl tRNA synthetases (Cusack *et al.*, 1996*a*,*b*; Goldgur *et al.*, 1997; Rould *et al.*, 1991), the effects of RNA's preference for A-form helices on recognition are immediately apparent. The proteins make numerous contacts in the shallow and exposed minor grooves of the RNA helices. This contrasts with the extensive use of the major groove in the recognition of B-form DNA helices. Beyond this generalization, the details of tRNA recognition differ in each specific case. Comparison of the protein-bound tRNA to the structure of free tRNA reveals that the proteins tend to distort the RNA conformation and partially unwind the helices near the anti-codon loop. In one case, namely the structure of glutamyl-tRNA synthetase (Rould *et al.*, 1991), the final base pair near the acceptor stem of the tRNA is broken, and the CCA acceptor makes a dramatic hairpin turn into the enzyme active site.

23.3.4.5. *Stem loops*

One fascinating observation in viewing the structures of RNA-binding proteins, even in the absence of RNA, is that aside from the tRNA-binding synthetases, they all appear to have evolved from or towards a very similar general fold (Burd & Dreyfuss, 1994). This fold, exemplified by the RNP domain found in numerous RNA-binding proteins, consists of a β-sheet surrounded on one side by α-helices and solvent-exposed on the opposing face. This general folding architecture is found in RNP domains, ribosome proteins, K-homologous domains (KH), double-stranded RNA-binding domains and cold shock proteins. Although each of these subsets of RNA-binding domains has a different topology and most probably bind to RNA with different surfaces, they all appear to have this alpha–beta–solvent architecture.

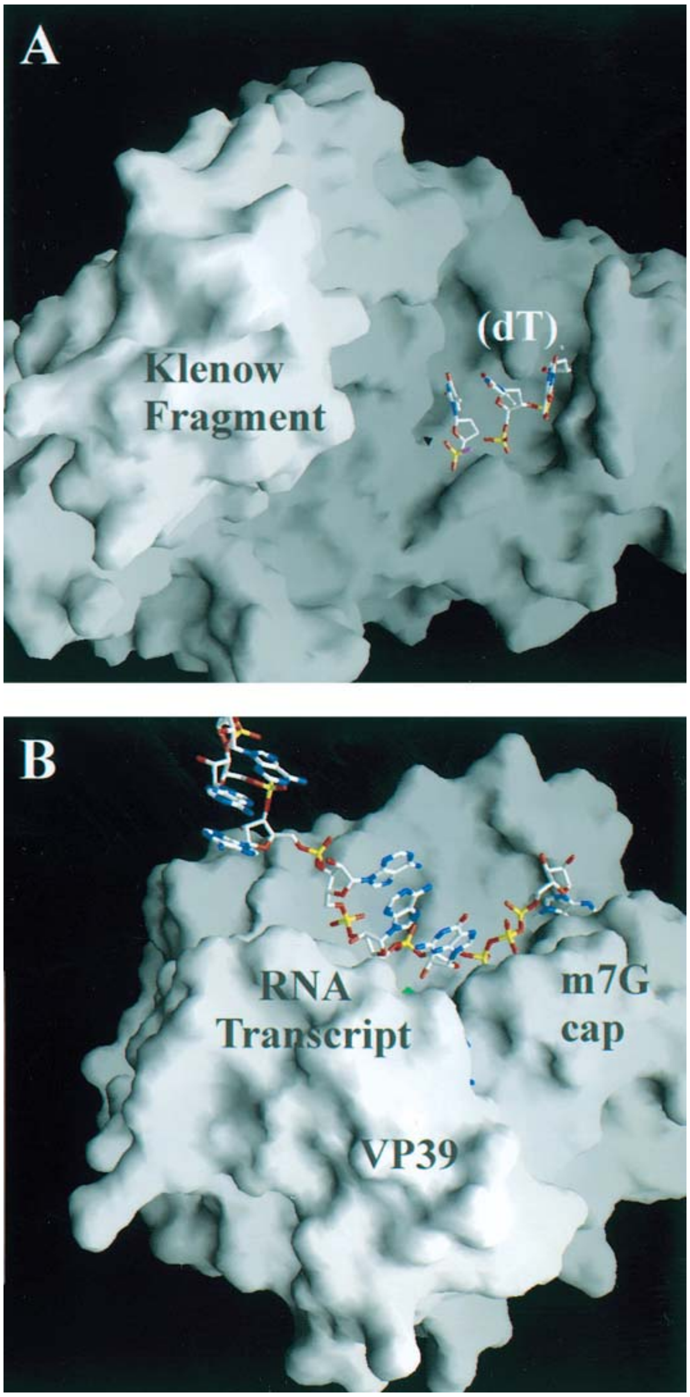


Figure 23.3.4.4
The sequence-nonspecific recognition of single-stranded nucleic acid. (*a*) Oligo(dT) bound in the exonuclease active site of DNA polymerase I Klenow fragment (Freemont *et al.*, 1988). (*b*) A short capped RNA transcript bound to the VP39 RNA methyltransferase (Hodel *et al.*, 1998). Both proteins primarily interact with the backbone of the nucleic acid.

Two proteins with this architecture have been co-crystallized with their specific RNA stem-loop ligands (Nagai *et al.*, 1995; van den Worm *et al.*, 1998). In both cases, the loop of the RNA binds to the open face of the β-sheet where solvent-exposed aromatic amino-acid side chains stack with the extrahelical bases of the RNA. Unpaired bases from the RNA also form numerous specific hydrogen bonds with protein side chains and polar backbone groups, imparting sequence specificity in the interaction. These structures suggest that the flat, open face of a β-sheet provides a good surface for RNA binding, where the extrahelical bases can make extensive and specific contacts with the protein.

23.3.4.6. Single-stranded sequence-nonspecific RNA–protein interactions

There is a single example of a single-stranded RNA–protein complex which is sequence-nonspecific. The structure of the vaccinia RNA methyltransferase VP39 bound to a 5′m^{7}G-capped RNA hexamer reveals a mechanism of nonspecific recognition reminiscent of the Klenow fragment–DNA tetramer complex (Hodel *et al.*, 1998). The RNA forms two short single-stranded helices of three bases each. The first of these helices binds in the active site of VP39 solely through hydrogen bonds between the protein and the ribose–phosphate backbone. The bases of the RNA strand stack together as trimers, but do not form any interactions with the protein (Fig. 23.3.4.4). Like the Klenow–DNA complex, this observation suggests an intuitive mechanism for sequence-nonspecific nucleic acid binding, where the single-stranded RNA forms short transient helices driven by intramolecular stacking interactions. The protein then recognizes and stabilizes the helical backbone conformation formed by this transient stacking without interacting with the bases themselves.

23.3.4.7. The recognition of alkylated bases

The complex of VP39 with capped RNA also illustrates a final example of the diversity of protein–ligand interactions in the specific recognition of the 7-methylguanosine cap. When guanosine is methylated at the N7 position, a positive charge is introduced to the π-ring system of the base. Eukaryotic cells utilize the methylation of a guanosine base at the N7 position as a tag or cap for the 5′ end of messenger RNA. The m^{7}G5′ppp mRNA cap is specifically recognized in the splicing of the first intron in nascent transcripts, in the transport of mRNA through the nuclear envelope and in the translation of the message by the ribosome (Varani, 1997). Two structures of specific m^{7}G binding proteins are now known: VP39 and the ribosomal cap-binding protein IF-4E (Hodel *et al.*, 1997; Marcotrigiano *et al.*, 1997). Each structure offers clues as to how the proteins can discriminate between the charged methylated m^{7}G base and the unmodified guanosine base. The m^{7}G base is stacked between aromatic protein side chains and hydrogen bonded to acidic protein residues (Fig. 23.3.4.5). One long-held hypothesis is that IF-4E, with dual tryptophan residues, binds specifically to the positively charged form of the base through a charge-transfer complex (Ueda, Iyo, Doi, Inoue & Ishida, 1991). The formation of a charge-transfer complex is evident in small-molecule studies and spectroscopic studies with IF-4E (Ueda, Iyo, Doi, Inoue, Ishida *et al.*, 1991). However, VP39 performs the same discrimination with the much less electronegative phenylalanine and tyrosine side chains (Hodel *et al.*, 1997). So far, no charge-transfer complex has been observed in VP39.

The recognition of charged methylated bases is important not only in mRNA processing, but also in the repair and recognition of DNA damaged by alkylating carcinogens. The mechanism by which the charged m^{7}G base is recognized is probably similar to how other positively charged bases, such as 3-methyladenosine, O2-methylcytosine and O2-methylthymidine, are recognized. In fact, the *E. coli* DNA repair enzyme, AlkA, will catalyse the glycolysis of all of these bases (Lindahl, 1982). The structure of AlkA is known, but only in the absence of a substrate (Labahn *et al.*, 1996). In this structure, a number of solvent-exposed tryptophan residues are found at the putative active site. This observation suggests that AlkA may recognize positively charged bases through an aromatic ‘sandwich’, much like that found in IF-4E and VP39.

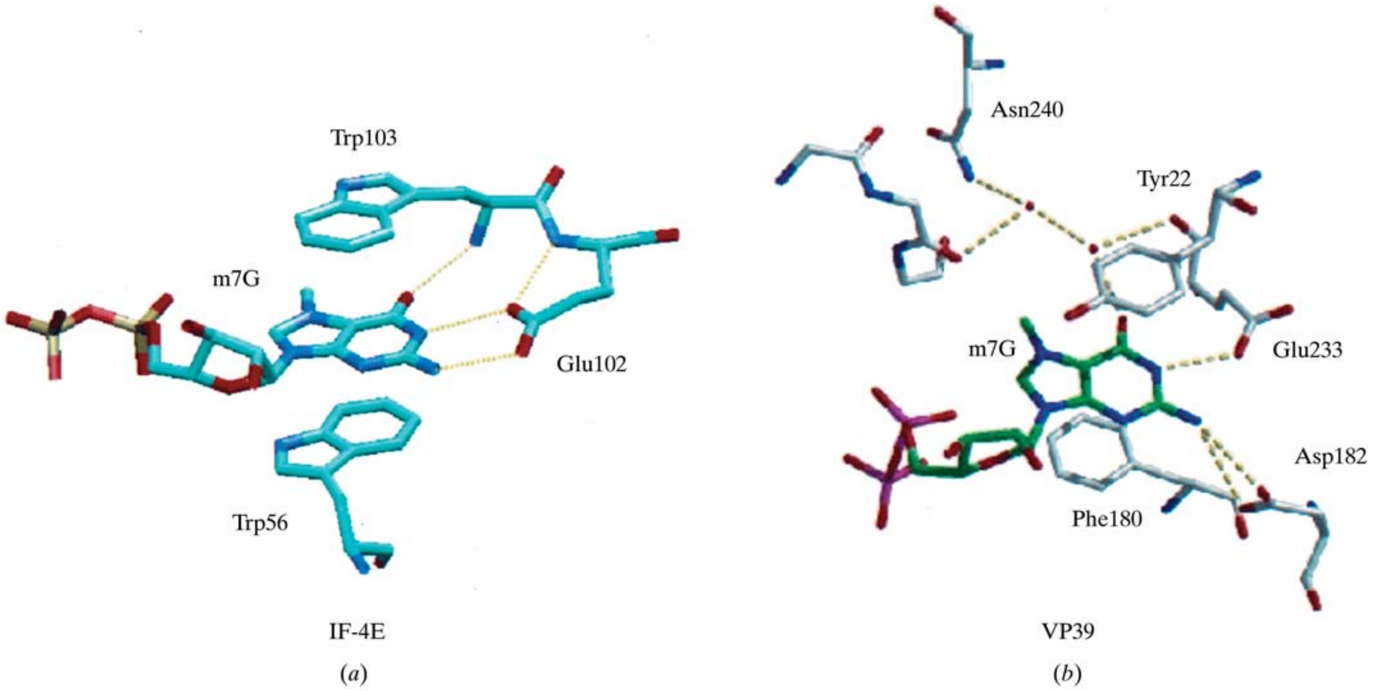


Figure 23.3.4.5
The specific recognition of the messenger RNA 7-methylguanosine cap. (*a*) The residues contacting the m^7G base in the cap-binding protein, IF-4E (Marcotrigiano *et al.*, 1997). (*b*) The residues interacting with the cap in the vaccinia RNA methyltransferase VP39 (Hodel *et al.*, 1997). Both proteins bind to the charged, methylated base by stacking aromatic amino acids on both sides of the base.

23.3.5. Phosphate and sulfate

Novel features of molecular recognition and electrostatic interactions of these two tetrahedral oxyanions have emerged from our crystallographic and functional studies of the phosphate-binding protein (PBP) and sulfate-binding protein (SBP), which serve as extremely specific initial receptors for ATP-binding cassette (ABC)-type active transport or permease in bacterial cells. The complexes of these proteins have K_d values in the low μM range. Although phosphate and sulfate are structurally similar, at physiological pH PBP and SBP exhibit no overlap in specificity (Medveczky & Rosenberg, 1971; Pardee, 1966; Jacobson & Quiocho, 1988). This stringent specificity prevents one tetrahedral oxyanion nutrient from becoming an inhibitor of transport for the other. The specificity of the PBP-dependent phosphate transport system is also shared by other phosphate transport systems in eukaryotic cells and across brush borders and into mitochondria.

As described below, discrimination between anions is based solely on the protonation state of the ligand. Sulfate, a conjugate base of a strong acid, is completely ionized at pH values above 3, whereas phosphate, a conjugate base of a weak acid, remains protonated up to pH 13.

The structure of the PBP–phosphate complex was initially determined at 1.7 Å resolution (Luecke & Quiocho, 1990). The resolution has been pushed to an ultra high resolution of 0.98 Å, the first reported for a protein with a molecular weight as high as 34 kDa with a bound ligand (Wang *et al.*, 1997). The bound phosphate is completely desolvated and sequestered in the protein cleft between two domains. It makes 12 hydrogen bonds with the proteins (11 with donor groups and one with an acceptor group), as well as one salt link to an Arg that is in turn salt-linked to an Asp residue (Fig. 23.3.5.1). The distances of the 12 hydrogen bonds between phosphate and PBP obtained from the ultra high resolution structure range from 2.432 to 2.906 Å (Wang *et al.*, 1997). The Asp56 carboxylate, the lone acceptor group, plays two key roles in conferring the exquisite specificity of PBP. It recognizes, by way of the hydrogen bond, a proton on the phosphate and presumably disallows, by charge repulsion, the binding of a fully ionized sulfate dianion (Luecke & Quiocho, 1990).

The SBP binding-site cleft is also tailor-made for sulfate (Pflugrath & Quiocho, 1985). In keeping with the stringent specificity of SBP for fully ionized tetrahedral oxyanions (Pardee, 1966; Jacobson & Quiocho, 1988), the bound sulfate, which is also completely dehydrated and buried, is held in place by seven hydrogen bonds made entirely with donor groups from uncharged polar residues of the protein (Fig. 23.3.5.2) (Pflugrath & Quiocho, 1985). The absence of a hydrogen-bond acceptor group accounts for the inability of SBP to bind phosphate. Interestingly, the absence of a salt link and the formation of five fewer hydrogen bonds with the bound sulfate (Fig. 23.3.5.2*b*) than with the bound phosphate (Fig. 23.3.5.1*b*) do not make the affinity of the SBP–sulfate complex any weaker than that of the PBP–phosphate complex. In fact, the sulfate binds 10–20 times more tightly to SBP (Pardee, 1966; Jacobson & Quiocho, 1988). Also, the hydration energies of both anions are likely to be similar.

The ability of PBP and SBP to differentiate each oxyanion ligand through the presence or absence of proton(s) is an extremely high level of sophistication in molecular recognition. The importance of complete hydrogen bonding in recognition of buried ligands is powerfully demonstrated in PBP and SBP. As the sulfate is fully ionized (*i.e.* possesses no hydrogen at physiological pH), repulsion occurs at Asp56 of PBP specifically for this dianion. On the other hand, SBP is unable to bind phosphate because it contains no hydrogen-bond acceptor in the binding site. Significantly, despite the potential for a large number of matched hydrogen-bonding pairs, a single mismatched hydrogen bond (*e.g.* a fully ionized sulfate providing no proton for interaction with Asp56 of PBP and no acceptor group in SBP for a phosphate proton) represents a binding energy barrier of 6–7 kcal mol^{-1} (1 kcal mol^{-1} = 4.184 kJ mol^{-1}).

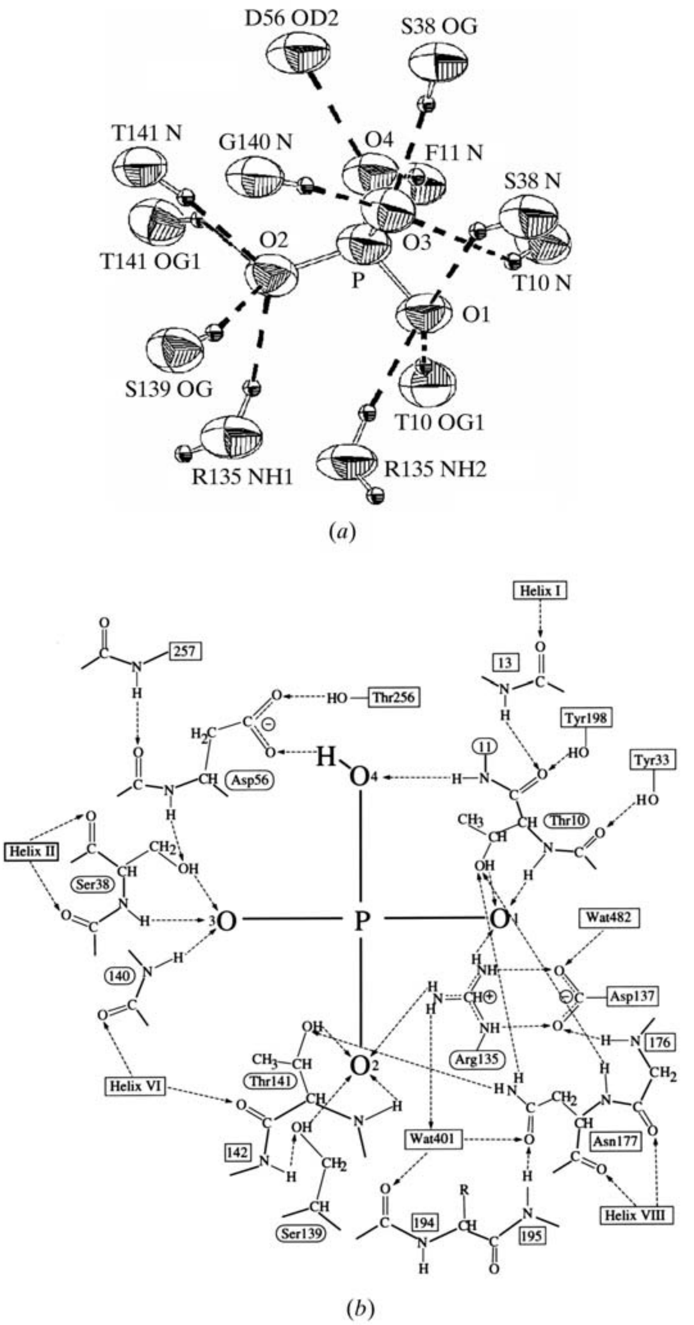

D56 OD2
S38 OG
T141 N
G140 N
O4
F11 N
S38 N
O2
O3
T141 OG1
P
T10 N
O1
S139 OG
T10 OG1
R135 NH1
R135 NH2
(a)
Helix I
257
13
Thr256
Tyr198
Asp56
Tyr33
Thr10
Helix II
Ser38
Wat482
140
Arg135
Asp137
176
Helix VI
Thr141
Wat401
Asn177
142
Ser139
194
195
Helix VIII
(b)

Figure 23.3.5.1
12 hydrogen-bonding interactions between the phosphate-binding protein (PBP) and phosphate. (*a*) Displacement ellipsoids of the atoms involved in the interactions from the 0.98 Å atomic structure (Wang *et al.*, 1997). (*b*) Schematic diagram of the interactions, including additional hydrogen bonds.

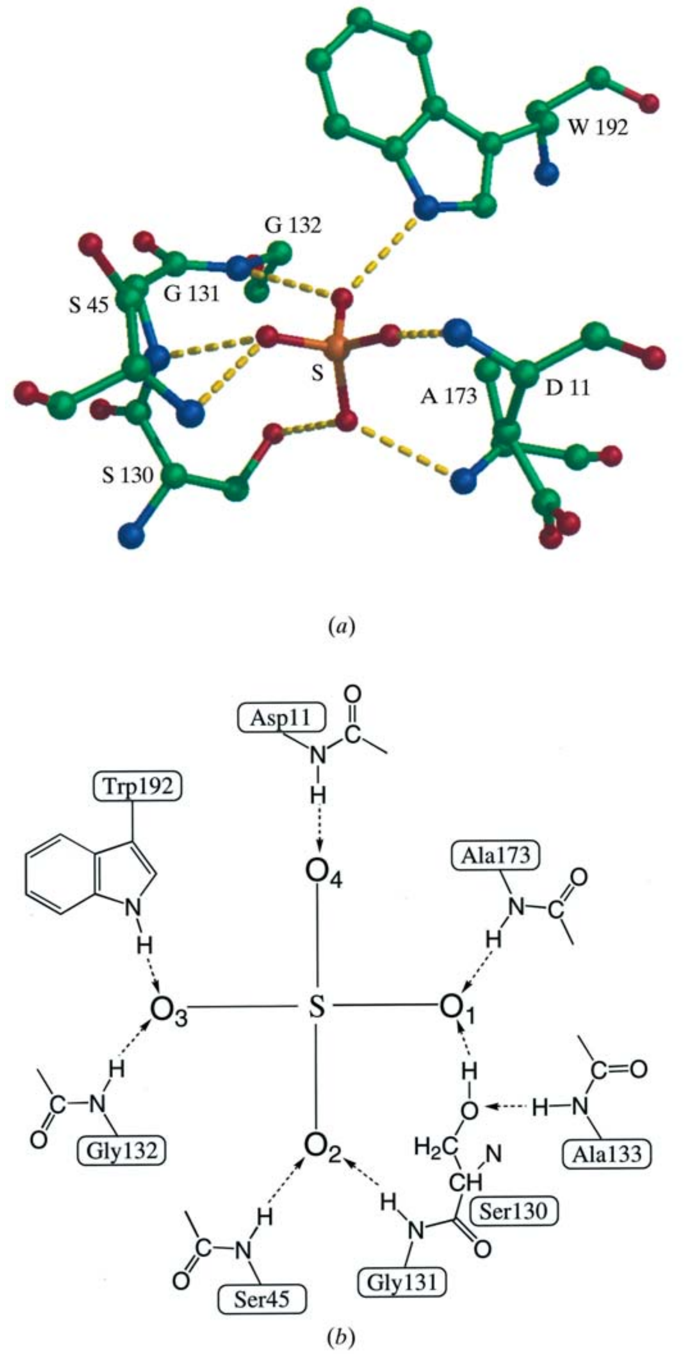
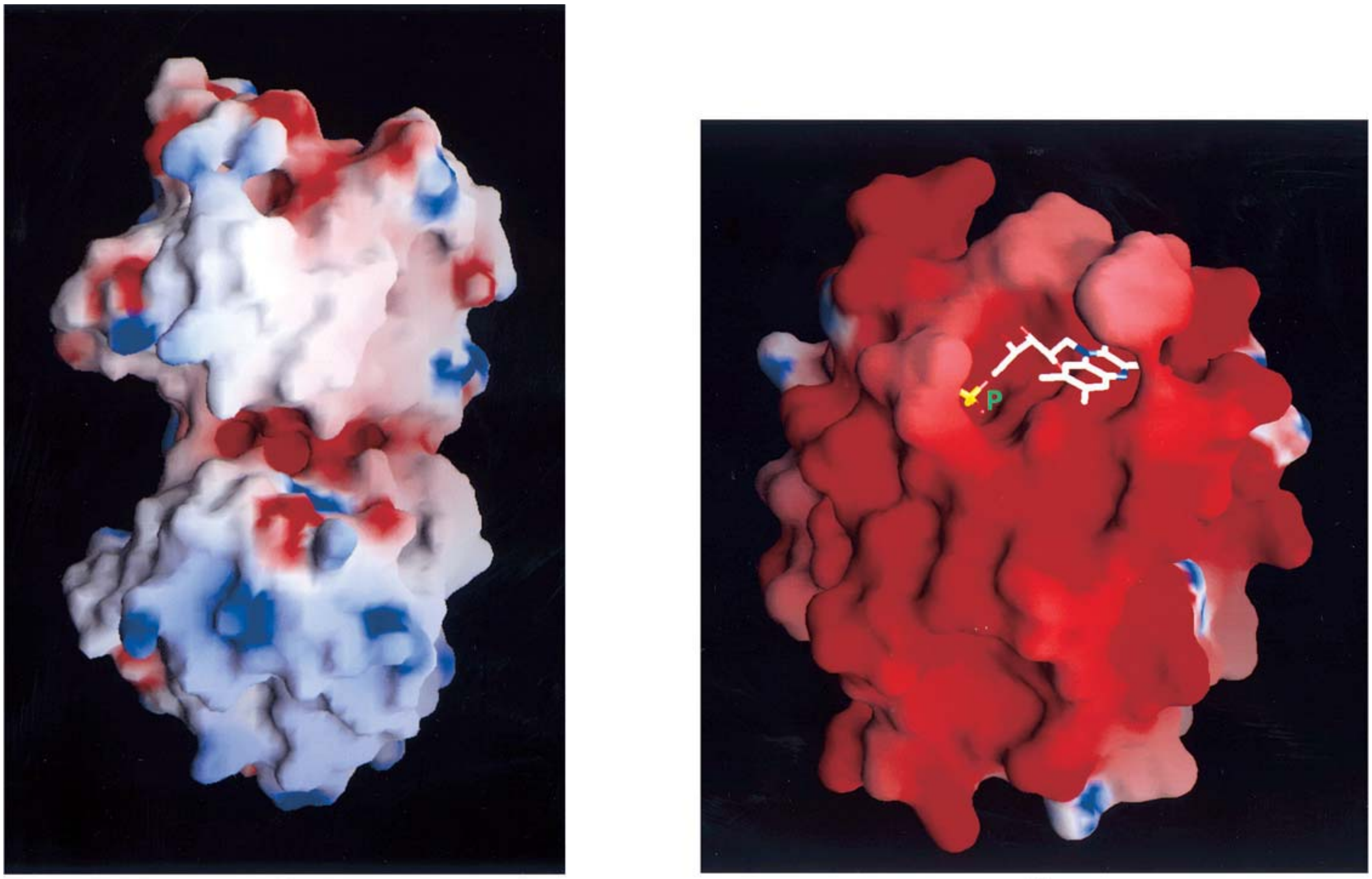

W 192
G 132
G 131
S 45
S
A 173
D 11
S 130
(a)
Asp11
Trp192
O4
Ala173
O3
S
O1
Gly132
O2
Ala133
Ser130
Ser45
Gly131
(b)

Figure 23.3.5.2
Seven hydrogen-bonding interactions between the sulfate-binding protein (SBP) and sulfate. (*a*) Interactions based on the 1.7 Å structure (J. Sack & F. A. Quiocho, unpublished data). (*b*) Schematic diagram of the interaction.

23.3.5.1. *Dominant role of local dipoles in stabilization of isolated charges*

A novel finding of further paramount importance and wide implication is how the isolated charges of the protein-bound phosphate and sulfate are stabilized. No counter-charged residues or cations are associated with the sulfate completely buried in SBP. Although a salt link involving Arg135 is formed with the phosphate bound to PBP, it is shared with an Asp residue (Fig. 23.3.5.1*b*). Moreover, site-directed mutagenesis studies indicate that phosphate binding is quite insensitive to modulation of the salt link (Yao *et al.*, 1996). These findings are a powerful demonstration of how a protein is able to stabilize the charges by means other than salt links. Experimental and computational studies indicate that local dipoles, including the hydrogen-bonding groups and the backbone NH groups from the first turn of helices, immediately surrounding the sulfate and phosphate are responsible for charge stabilization (Pflugrath & Quiocho, 1985; Quiocho *et al.*, 1987; Åqvist *et al.*, 1991; He & Quiocho, 1993; Yao *et al.*, 1996; Ledvina *et al.*, 1996). Helix macrodipoles play little or no role in charge stabilization of the anions. The same principle of charge stabilization by local dipoles also applies for the following buried uncompensated ionic groups: Arg151 of the arabinose-binding protein (Quiocho *et al.*, 1987), the zwitterionic leucine ligand bound to the leucine/isoleucine/valine-binding protein (Quiocho *et al.*, 1987), the potassium in the pore of the potassium channel (Doyle *et al.*, 1998) and Arg56 of synaptobrevin-II in a SNARE complex (Sutton *et al.*, 1998).



(*a*) (*b*)

Figure 23.3.5.3
Electrostatic surface potential of (*a*) the phosphate-binding protein and (*b*) flavodoxin. The molecular surface electrostatic potentials, calculated and displayed using *GRASP* (Nicholls *et al.*, 1991), are −10 kT (red), neutral (white) and +10 kT (blue) [see Ledvina *et al.* (1996) for more details]. (*a*) Wild-type phosphate-binding protein based on the X-ray structure of the open cleft, unliganded form (Ledvina *et al.*, 1996). The phosphate-binding site is located in the cleft (with negative surface potential) in the middle of the molecule and between the two domains. (*b*) Flavodoxin with bound flavin mononucleotide (FMN). The phosphoryl group (P) of the FMN is bound in a pocket with intense negatively charge surface potential. The surface potential was calculated without the bound flavin mononucleotide using the structure from the Protein Data Bank (PDB code: 2fox).

23.3.5.2. Short hydrogen bonds

The ultra high resolution refined structure of the PBP–phosphate complex is the first to show structurally the formation of an extremely short hydrogen bond (2.432 Å) between the Asp56 carboxylate of PBP and phosphate. Although this short hydrogen bond is within the proposed range of low-barrier hydrogen bonds with estimated energies of 12–24 kcal mol^{-1} (Hibbert & Emsley, 1990), its contribution to phosphate binding affinity has been assessed to be no better than that of a normal hydrogen bond (Wang *et al.*, 1997). Thus, a unique role for short hydrogen bonds in biological systems, such as in enzyme catalysis (Gerlt & Gassman, 1993; Cleland & Kreevoy, 1994), remains controversial.

23.3.5.3. Non-complementary negative electrostatic surface potential of protein sites specific for anions

The presence of an uncompensated negatively charged Asp56 is unusual for an anion-binding site, as observed in PBP. In fact, a related discovery of profound ramification is that the binding-cleft region of PBP has an intense negative electrostatic surface potential (Fig. 23.3.5.3*a*) (Ledvina *et al.*, 1996). Non-complementarity between the surface potential of a binding region and an anion ligand is not unique to PBP. We have reported similar findings for SBP, a DNA-binding protein, and, even more dramatically, for the redox protein flavodoxin (Fig. 23.3.5.3*b*) (Ledvina *et al.*, 1996). Evidently, for proteins such as these, which rely on hydrogen-bonding interactions with only uncharged polar residues for anion binding and electrostatic balance, a non-complementary surface potential is not a barrier to binding. This conclusion is supported by very recent fast kinetic studies of binding of phosphate to PBP and the effect of ionic strength on binding (Ledvina *et al.*, 1998).

FAQ is an HHMI Investigator. The work carried out in his laboratory is supported in part by grants from NIH and the Welch Foundation.

References

Åqvist, J., Luecke, H., Quiocho, F. A. & Warshel, A. (1991). *Dipoles localized at helix termini of proteins stabilize charges*. *Proc. Natl Acad. Sci. USA*, **88**, 2026–2030.

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *On the conformation of the hen egg-white lysozyme molecule*. *Proc. R. Soc. London Ser. B Biol. Sci.* **167**, 365–377.

Bochkarev, A., Pfuetzner, R. A., Edwards, A. M. & Frappier, L. (1997). *Structure of the single-stranded-DNA-binding domain of replication protein A bound to DNA*. *Nature (London)*, **385**, 176–181.

Burd, C. G. & Dreyfuss, G. (1994). *Conserved structures and diversity of functions of RNA-binding proteins*. *Science*, **265**, 615–621.

Cheng, X. (1995). *DNA modification by methyltransferases*. *Curr. Opin. Struct. Biol.* **5**, 4–10.

Cleland, W. W. & Kreevoy, M. M. (1994). *Low-barrier hydrogen bonds and enzymic catalysis*. *Science*, **264**, 1887–1890.

Cotton, F. A., Hazen, E. E. Jr & Legg, M. J. (1979). *Staphylococcal nuclease: proposed mechanism of action based on structure of enzyme–thymidine 3′,5′-bisphosphate–calcium ion complex at 1.5-Å resolution*. *Proc. Natl Acad. Sci. USA*, **76**, 2551–2555.

Cusack, S., Yaremchuk, A. & Tukalo, M. (1996*a*). *The crystal structure of the ternary complex of T. thermophilus seryl-tRNA synthetase with tRNA(Ser) and a seryl-adenylate analogue reveals a conformational switch in the active site*. *EMBO J.* **15**, 2834–2842.

Cusack, S., Yaremchuk, A. & Tukalo, M. (1996*b*). *The crystal structures of T. thermophilus lysyl-tRNA synthetase complexed with E. coli tRNA(Lys) and a T. thermophilus tRNA(Lys) transcript: anticodon recognition and conformational changes upon binding of a lysyl-adenylate analogue*. *EMBO J.* **15**, 6321–6334.

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A. L., Gulbis, J. M., Cohen, S. L., Chait, B. T. & MacKinnon, R. (1998). *Science*, **280**, 68–77.

Freemont, P. S., Friedman, J. M., Beese, L. S., Sanderson, M. R. & Steitz, T. A. (1988). *Cocrystal structure of an editing complex of Klenow fragment with DNA*. *Proc. Natl Acad. Sci. USA*, **85**, 8924–8928.

Gerlt, J. A. & Gassman, P. G. (1993). *Understanding the rates of certain enzyme-catalyzed reactions: proton abstraction from carbon acids, acyl-transfer reaction, and displacement reactions of phosphodiesters*. *Biochemistry*, **32**, 1943–1952.

Glusker, J. P. (1991). *Structural aspects of metal liganding to functional groups in proteins*. *Adv. Protein Chem.* **42**, 1–76.

Goldgur, Y., Mosyak, L., Reshetnikova, L., Ankilova, V., Lavrik, O., Khodyreva, S. & Safro, M. (1997). *The crystal structure of phenylalanyl-tRNA synthetase from Thermus Thermophilus complexed with cognate tRNAPhe*. *Structure*, **5**, 59–68.

Harrington, R. E. & Winicov, I. (1994). *New concepts in protein–DNA recognition: sequence-directed DNA bending and flexibility*. *Prog. Nucleic Acid Res. Mol. Biol.* **47**, 195–270.

He, J. J. & Quiocho, F. A. (1993). *Dominant role of local dipoles in stabilizing uncompensated charges on a sulphate sequestered in a periplasmic active transport*. *Protein Sci.* **2**, 1643–1647.

Hibbert, F. & Emsley, J. (1990). *Hydrogen bonding and chemical reactivity*. *Adv. Phys. Org. Chem.* **226**, 255–379.

Hodel, A. E., Gershon, P. D. & Quiocho, F. A. (1998). *Structural basis for sequence non-specific recognition of 5′-capped mRNA by a cap-modifying enzyme*. *Mol. Cell*, **1**, 443–447.

Hodel, A. E., Gershon, P. D., Shi, X., Wang, S. M. & Quiocho, F. A. (1997). *Specific protein recognition of an mRNA cap through its alkylated base*. (Letter.) *Nature Struct. Biol.* **4**, 350–354.

Jacobson, B. L. & Quiocho, F. A. (1988). *Sulphate-binding protein dislikes protonated oxyacids: a molecular explanation*. *J. Mol. Biol.* **204**, 783–787.

Joachimiak, A., Schevitz, R. W., Kelley, R. L., Yanofsky, C. & Sigler, P. B. (1983). *Functional inferences from crystals of Escherichia coli trp repressor*. *J. Biol. Chem.* **258**, 12641–12643.

Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. (1990). *Crystal structure of an engrailed homeodomain–DNA complex at 2.8 Å resolution: a framework for understanding homeodomain–DNA interactions*. *Cell*, **63**, 579–590.

Labahn, J., Scharer, O. D., Long, A., Ezaz-Nikpay, K., Verdine, G. L. & Ellenberger, T. E. (1996). *Structural basis for the excision repair of alkylation-damaged DNA*. *Cell*, **86**, 321–329.

Ledvina, P. S., Tsai, A.-H., Wang, Z., Koehl, E. & Quiocho, F. A. (1998). *Dominant role of local dipolar interactions in phosphate binding to a receptor cleft with an electronegative charge surface potential: equilibrium, kinetic and crystallographic studies*. *Protein Sci.* **7**, 2550–2559.

Ledvina, P. S., Yao, N., Choudhary, A. & Quiocho, F. A. (1996). *Negative electrostatic surface potential of protein sites specific for anionic ligands*. *Proc. Natl Acad. Sci. USA*, **93**, 6786–6791.

Lindahl, T. (1982). *DNA repair enzymes*. *Annu. Rev. Biochem.* **51**, 61–87.

Luecke, H. & Quiocho, F. A. (1990). *High specificity of a phosphate transport protein determined by hydrogen bonds*. *Nature (London)*, **347**, 402–406.

Marcotrigiano, J., Gingras, A. C., Sonenberg, N. & Burley, S. K. (1997). *X-ray studies of the messenger RNA 5′ cap-binding protein (eIF4E) bound to 7-methyl-GDP*. *Nucleic Acids Symp. Ser.* pp. 8–11.

Meador, W. E., George, S. E., Means, A. R. & Quiocho, F. A. (1995). *X-ray analysis reveals conformational adaptation of the linker in functional calmodulin mutants*. (Letter.) *Nature Struct. Biol.* **2**, 943–945.

Medveczky, N. & Rosenberg, H. (1971). *Phosphate transport in Escherichia coli*. *Biochim. Biophys. Acta*, **241**, 494–506.

Nagai, K. (1996). *RNA–protein complexes*. *Curr. Opin. Struct. Biol.* **6**, 53–61.

Nagai, K., Oubridge, C., Ito, N., Jessen, T. H., Avis, J. & Evans, P. (1995). *Crystal structure of the U1A spliceosomal protein complexed with its cognate RNA hairpin*. *Nucleic Acids Symp. Ser.* 1–2.

Nicholls, A., Sharp, K. & Honig, B. (1991). *Protein folding and association; insights from the interfacial and thermodynamic properties of hydrocarbons*. *Proteins*, **11**, 281–296.

Orgel, L. E. (1966). *An Introduction to Transition-Metal Chemistry. Ligand-Field Theory*, 2nd ed. London: Methuen and New York: Wiley.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988*a*). *Crystal structure of trp repressor/operator complex at atomic resolution*. *Nature (London)*, **335**, 321–329.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988*b*). *Crystal structure of trp repressor/operator complex at atomic resolution. Erratum*. *Nature (London)*, **335**, 837.

Pabo, C. O. & Sauer, R. T. (1992). *Transcription factors: structural families and principles of DNA recognition*. *Annu. Rev. Biochem.* **61**, 1053–1095.

Pardee, A. B. (1966). *Purification and properties of a sulfate-binding protein from Salmonella typhimurium*. *J. Biol. Chem.* **241**, 5886–5892.

Pavletich, N. P. & Pabo, C. O. (1991). *Zinc finger–DNA recognition: crystal structure of a Zif268–DNA complex at 2.1 Å*. *Science*, **252**, 809–817.

Pflugrath, J. W. & Quiocho, F. A. (1985). *Sulphate sequestered in the sulphate-binding protein of Salmonella typhimurium is bound solely by hydrogen bonds*. *Nature (London)*, **314**, 257–260.

Quiocho, F. A. (1986). *Carbohydrate-binding proteins: tertiary structures and protein–sugar interactions*. *Annu. Rev. Biochem.* **55**, 287–315.

Quiocho, F. A., Sack, J. S. & Vyas, N. K. (1989). *Substrate specificity and affinity of a protein modulated by bound water molecules*. *Nature (London)*, **340**, 404–407.

Quiocho, F. A., Wilson, D. K. & Vyas, N. K. (1987). *Stabilization of charges on isolated charged groups sequestered in proteins by polarized peptide units*. *Nature (London)*, **329**, 561–564.

Rademacher, T. W., Parekh, R. B. & Dwek, R. A. (1988). *Glycobiology*. *Annu. Rev. Biochem.* **57**, 785–838.

Rould, M. A., Perona, J. J. & Steitz, T. A. (1991). *Structural basis of anticodon loop recognition by glutaminyl-tRNA synthetase*. *Nature (London)*, **352**, 213–218.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). *Sequence-specific recognition of double helical nucleic acids by proteins*. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.

Shimon, L. J. & Harrison, S. C. (1993). *The phage 434 OR2/R1–69 complex at 2.5 Å resolution*. *J. Mol. Biol.* **232**, 826–838.

Steitz, T. A. (1990). *Structural studies of protein–nucleic acid interaction: the sources of sequence-specific binding*. *Q. Rev. Biophys.* **23**, 205–280.

Sutton, R. B., Fasshauer, D., Jan, R. & Brunger, A. T. (1998). *Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution*. *Nature (London)*, **395**, 347–353.

Tucker, P. W., Hazen, E. E. Jr & Cotton, F. A. (1979). *Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. III. Correlation of the three-dimensional structure with the mechanisms of enzymatic action*. *Mol. Cell. Biochem.* **23**, 67–86.

Ueda, H., Iyo, H., Doi, M., Inoue, M. & Ishida, T. (1991). *Cooperative stacking and hydrogen bond pairing interactions of fragment peptide in cap binding protein with mRNA cap structure*. *Biochim. Biophys. Acta*, **1075**, 181–186.

Ueda, H., Iyo, H., Doi, M., Inoue, M., Ishida, T., Morioka, H., Tanaka, T., Nishikawa, S. & Uesugi, S. (1991). *Combination of Trp and Glu residues for recognition of mRNA cap structure. Analysis of m7G base recognition site of human cap binding protein (IF-4E) by site-directed mutagenesis*. *FEBS Lett.* **280**, 207–210.

Varani, G. (1997). *A cap for all occasions*. *Structure*, **5**, 855–858.

Volbeda, A., Fontecilla-Camps, J. C. & Frey, M. (1996). *Novel metal sites in protein structures*. *Curr. Opin. Struct. Biol.* **6**, 804–812.

Vyas, N. K. (1991). *Atomic features of protein–carbohydrate interactions*. *Curr. Opin. Struct. Biol.* **1**, 732–740.

Vyas, N. K., Vyas, M. N. & Quiocho, F. A. (1988). *Sugar and signal-transducer binding sites of the Escherichia coli galactose chemoreceptor protein. Science*, **242**, 1290–1295.

Wang, Z., Luecke, H., Yao, N. & Quiocho, F. A. (1997). *Nature Struct. Biol.* **4**, 519–522.

Weis, W. I. & Drickamer, K. (1996). *Structural basis of lectin–carbohydrate recognition. Annu. Rev. Biochem.* **65**, 441–473.

Werner, M. H., Gronenborn, A. M. & Clore, G. M. (1996*a*). *Intercalation, DNA kinking, and the control of transcription. Science*, **271**, 778–784.

Werner, M. H., Gronenborn, A. M. & Clore, G. M. (1996*b*). *Intercalation, DNA kinking, and the control of transcription. Erratum. Science*, **272**, 19.

Worm, S. H. van den, Stonehouse, N. J., Valegard, K., Murray, J. B., Walton, C., Fridborg, K., Stockley, P. G. & Liljas, L. (1998). *Crystal structures of MS2 coat protein mutants in complex with wild-type RNA operator fragments*. *Nucleic Acids Res.* **26**, 1345–1351.

Yao, N., Ledvina, P. S., Choudhary, A. & Quiocho, F. A. (1996). *Modulation of a salt link does not affect binding of phosphate to its specific active transport receptor. Biochemistry*, **35**, 2079–2085.

International Tables for Crystallography (2012). Vol. F, Chapter 23.4, pp. 766–799.

Chapter 23.4. Nucleic acids

R. E. Dickerson

23.4.1. Introduction

In 1953, James Watson and Francis Crick solved the structure of double-helical DNA (Watson & Crick, 1953; Crick & Watson, 1954). So what has a dedicated cadre of X-ray crystallographers been doing for the subsequent 45 years? That is the subject of this chapter: the advance of our knowledge of nucleic acid duplexes, primarily from single-crystal X-ray diffraction, and the biological implications of this new knowledge. The focus will be primarily on DNA because much more is known about it, but DNA/RNA hybrids and duplex RNA will also be considered. Because the emphasis is on the geometry of the nucleic acid double helix, exotic structures, such as quadruplexes, hammerhead ribozymes and aptamers, will be omitted, as will larger-scale structures such as tRNA.

Fibre diffraction showed that there were two basic forms of DNA duplex: the common B form and a more highly crystalline A form (Fig. 23.4.1.1) that, in some but not all sequences, could be produced by dehydrating the fibre (Franklin & Gosling, 1953; Langridge *et al.*, 1960; Arnott, 1970; Leslie *et al.*, 1980). A- and B-DNA are contrasted in Figs. 23.4.1.2 and 23.4.1.3. The high-humidity B form has base pairs sitting squarely on the helix axis and roughly perpendicular to that axis. In contrast, in the low-humidity A form, the base pairs are displaced off the helix axis by *ca* 4 Å and are inclined 10–20° away from perpendicularity to that axis. The two grooves in B-DNA are of comparable depth because base pairs sit *on* the helix axis, but the major groove is wider than the minor because of asymmetry of attachment of base pairs to the backbone chains. In A-DNA, the minor groove is broad and shallow, whereas the major groove is cavernously deep (all the way from the surface of the helix, to the helix axis, and beyond) but can be quite narrow.

Pohl and co-workers had shown in the 1970s that alternating poly(dC-dG) is special in that it undergoes a reversible salt- or alcohol-induced conformation change (Pohl & Jovin, 1972; Pohl, 1976). Hence, it was not surprising that when DNA synthesis methods advanced to the stage where oligonucleotide crystallization became feasible, two separate research groups – those of Alexander Rich at MIT and Richard Dickerson at Caltech – elected to synthesize, crystallize and solve a short, alternating C-G oligomer. The result was a third family of DNA duplexes, Z-DNA (Fig. 23.4.1.4), first as the hexamer C-G-C-G-C-G (Z1) and then the tetramer C-G-C-G (Z3). (References to A-, B- and Z-DNA structures are listed at the end of Tables A23.4.1.1, A23.4.1.2 and A23.4.1.3 in the Appendix, respectively. They are cited by numbers beginning with A, B or Z.) Single-crystal analyses of the traditional helix types soon followed: B-DNA as C-G-C-G-A-A-T-T-C-G-C-G (B1), and A-DNA as both C-C-G-G (A1) and G-G-T-A-T-A-C-C (A2).

23.4.2. Helix parameters

23.4.2.1. *Backbone geometry*

Before making detailed comparisons of the three helix types, one must define the parameters by which the helices are characterized. The fundamental feature of all varieties of nucleic acid double helices is two antiparallel sugar–phosphate backbone chains, bridged by paired bases like rungs in a ladder (Fig. 23.4.2.1). Using the convention that the positive direction of a backbone chain is from 5′ to 3′ within a nucleotide, the right-hand chain in Fig. 23.4.2.1 runs downward, while the left-hand chain runs upward. A- or B-DNA is then obtained by twisting the ladder into a right-handed helix. But Z-DNA cannot be obtained from Fig. 23.4.2.1 simply by giving it a left-handed twist; both backbone chains run in the wrong direction for Z-DNA. A more complex adjustment is required, and this will be addressed again later.

The conformation of the backbone chain along each nucleotide is described by six torsion angles, labelled α through ζ, as shown in Fig. 23.4.2.2. An earlier convention termed these same six angles as ω, φ, ψ, ψ', φ', ω' (Sundaralingam, 1975), but the alphabetical nomenclature is now generally employed. Torsion angles are defined in Fig. 23.4.2.3, which also shows three common configurations: *gauche*$^-$ (−60°), *trans* (180°) and *gauche*$^+$ (+60°). These three configurations are especially favoured with sp^3 hybridization or tetrahedral ligand geometry at the two ends of the bond in question, because their 'staggered' arrangement minimizes ligand–ligand interactions across the bond. An 'eclipsed' arrangement with ligands at −120°, 0° (*cis*), and 120° is unfavourable because it brings substituents at the two ends of the bond into opposition. Table 23.4.2.1 lists the mean values and standard deviations of all six main-chain torsion angles for A-, B- and Z-DNA, as recently observed in 96 oligonucleotide crystal structures (Schneider *et al.*, 1997).

23.4.2.2. *Sugar ring conformations*

The type of ligand–ligand clash just mentioned is an important element in ensuring that five-membered rings, such as ribose and deoxyribose, are not ordinarily planar, even though the internal bond angle of a regular pentagon, 108°, is close to the 109.5° of tetrahedral geometry. A stable compromise is for one of the four ring atoms to lie out of the plane defined by the other four, as in Fig. 23.4.2.4. This is termed an 'envelope' or E conformation, by

---

This chapter is dedicated to Irving Geis, who died on 22 July 1997 at the age of 88, just as the chapter was begun. Irv was a pioneer in the representation of protein and DNA structures, beginning with illustrations for *Scientific American* articles on myoglobin (Kendrew, 1961), lysozyme (Phillips, 1966), cytochrome *c* (Dickerson, 1972) and DNA (Dickerson, 1983). He was coauthor with the present writer of *Structure and Action of Proteins* (Dickerson & Geis, 1969) and two later textbooks (Dickerson & Geis, 1976, 1983) and contributed drawings and paintings to a great number of other books and articles, most notably Voet & Voet's *Biochemistry* (Voet & Voet, 1990, 1995), which is a veritable gallery of Irv's art. His meticulous and carefully thought-out diagrams and drawings of myoglobin and haemoglobins have never been matched. More information about his life, work and art may be found in three articles by the present author (Dickerson, 1997*a*,*b*,*c*). Irv saw his role as one of bringing an understanding of protein structure to life scientists and sometimes referred to himself half-humorously as 'the Andreas Vesalius of molecular anatomy'. In view of the formative influence that his art exerted on the first generation of protein crystallographers and molecular biologists, it is more appropriate to remember Irv as the Leonardo da Vinci of macromolecules. As of late 2000, nearly all of Irving Geis' work – paintings, drawings, illustrations and correspondence – is being preserved for study as the Geis Archives at the Howard Hughes Medical Institute, Washington DC.

Copyright © 2012 International Union of Crystallography

analogy with a four-cornered envelope having a flap at an angle. Intermediate ‘twist’ or T forms are also possible, in which two adjacent atoms sit on either side of the plane defined by the other three, but this discussion will focus on the simple envelope conformations. In most cases, the accuracy of a nucleic acid crystal structure determination is such that it would be difficult to distinguish clearly between a given E form and its flanking T forms. For this reason, most structure reports consider only the E alternatives.

A convenient and intuitive nomenclature is to name the conformation after the out-of-plane atom and then specify whether it is out of plane on the same side as the C5′ atom (*endo*) or the opposite side (*exo*). Ten such conformations exist: five *endo* and five *exo*. In Fig. 23.4.2.4 (top), pushing the C3′ atom of the C3′-*endo* conformation into the plane of the ring would tend to push C2′ below the ring, passing through a T state and creating a C2′-*exo* conformation. C2′ can, in turn, be returned to the ring plane if C1′ is pushed above the ring, forming C1′-*endo*, and so on, around the ring. In this way, a contiguous series of alternating *endo/exo* conformations is produced, as listed in Table 23.4.2.2.

This ten-conformation *endo/exo* cycle can be generalized to a continuous distribution of intermediate conformations, characterized by a pseudorotation angle, P (Altona *et al.*, 1968; Altona & Sundaralingam, 1972), with the ten *endo/exo* conformations spaced 36° apart (Table 23.4.2.2). Fig. 23.4.2.5 shows the calculated potential energy of conformations around the pseudorotation cycle (Levitt & Warshel, 1978). Note that C2′-*endo* and C3′-*endo* are most stable, that the pathway between them along the right half of the circle remains one of low energy, but that a large $6\,\text{kcal mol}^{-1}$ potential energy barrier ($1\,\text{kcal mol}^{-1} = 4.184\,\text{kJ mol}^{-1}$) effectively forbids conformations around the left half of the circle.

As Fig. 23.4.2.4 indicates, the main-chain torsion angle, δ, is sensitive to ring conformation, because the C5′—C4′ and C3′—O3′ bonds that define the angle shift as ring puckering changes. The idealized relationship between torsion angle, δ, and pseudorotation angle, P (Saenger, 1984), is

$$\delta = 40^\circ \cos(P + 144^\circ) + 120^\circ.$$

Fig. 23.4.2.6 shows the observed torsion angles, δ, and pseudorotation angles, P, from X-ray crystal structure analyses of synthetic DNA oligonucleotides: 296 examples from A-DNA and 280 from B-DNA. The most striking aspect of this plot is the radically different behaviour of A- and B-DNA. The prototypical sugar conformation for A-DNA obtained from fibre diffraction modelling, C3′-*endo*, is, in fact, adhered to quite closely in A-DNA crystal structures.

However, B-DNA shows a quite different behaviour. Although earlier fibre diffraction led one to expect C2′-*endo* sugars, the actual experimental distribution is quite broad, extending up the right-hand side of the pseudorotation circle of Fig. 23.4.2.5, through C1′-*exo*, O1′-*endo* and C4′-*exo*, in some cases all the way to C3′-*endo* itself. Indeed, the mean value of δ observed in B-DNA oligomer crystal structures is 128° rather than 144° (Table 23.4.2.1), making C1′-*exo* a better description of sugar conformation in B-DNA than C2′-*endo*. Old habits die hard, however, and the B-DNA sugar conformation is still colloquially termed C2′-*endo*, a designation of historical significance but of little practical value. The apparent greater malleability of the B helix compared to A may indeed be one feature that makes B-DNA particularly suitable for expressing its base sequence to drugs and control proteins *via* local helix structure changes.

23.4.2.3. Base pairing

The key to the biological role of DNA is that one of the two purines can pair with only one of the pyrimidines: A with

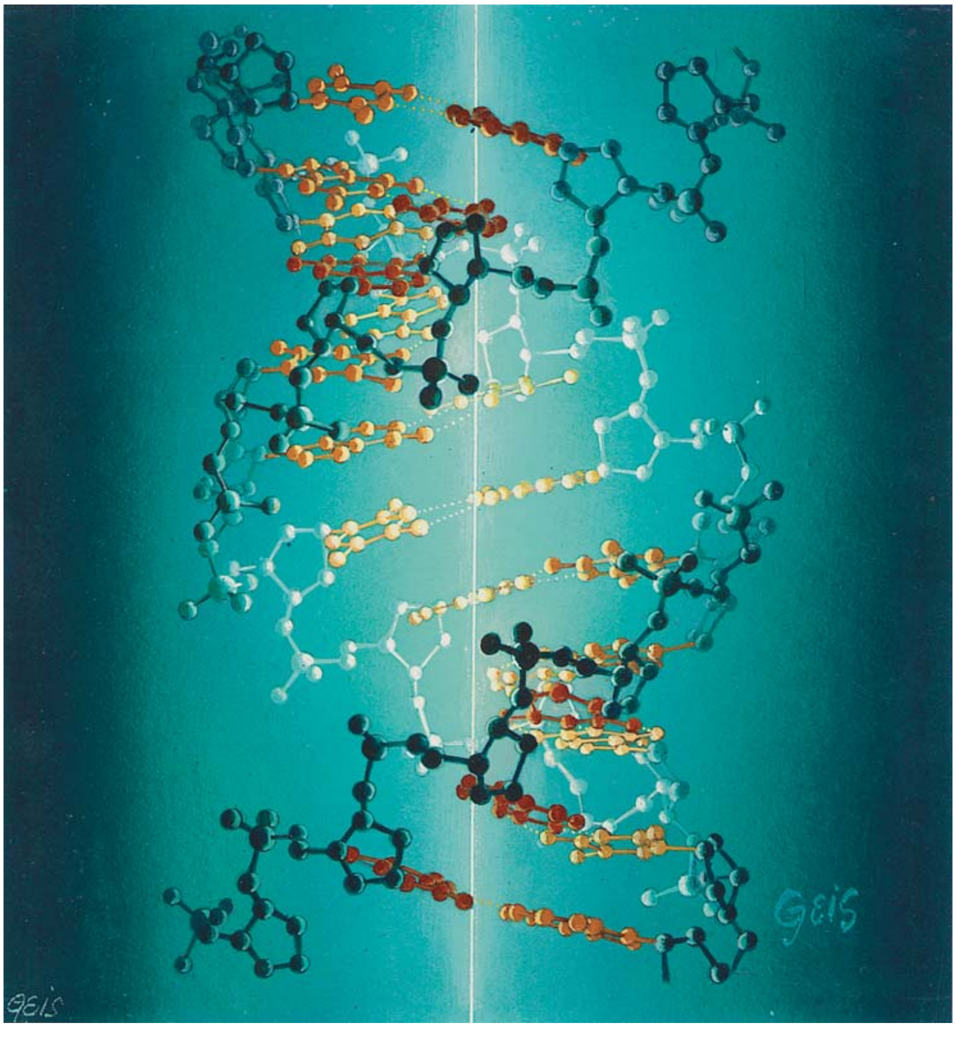


Figure 23.4.1.1
‘Hot wire’ painting of A-DNA by Irving Geis. Geis produced two dramatic paintings of horse-heart cytochrome *c*, in which the sole light source was the central iron atom within the haem, producing a glowing ‘molecular lantern’ effect. One painting showed this central luminous haem surrounded by hydrophobic side chains; the other featured the polar side chains extending out from the surface. These are to be seen today on the front and back covers of Voet & Voet’s *Biochemistry* (Voet & Voet, 1990, 1995). In the present A-DNA painting, Geis chose the imaginary central axis of the helix as a monofilament light source, thereby reversing the conventional illumination: atoms lining the deep major groove glow brightly, whereas the outer surface of the helix is in dark silhouette. Geis struggled with the B helix as an artistic subject, but was never satisfied with the results. Hence, this glowing A-DNA helix represents his nucleic acid artistic legacy. Reprinted courtesy of the estate of Irving Geis. Rights owned by Howard Hughes Medical Institute.

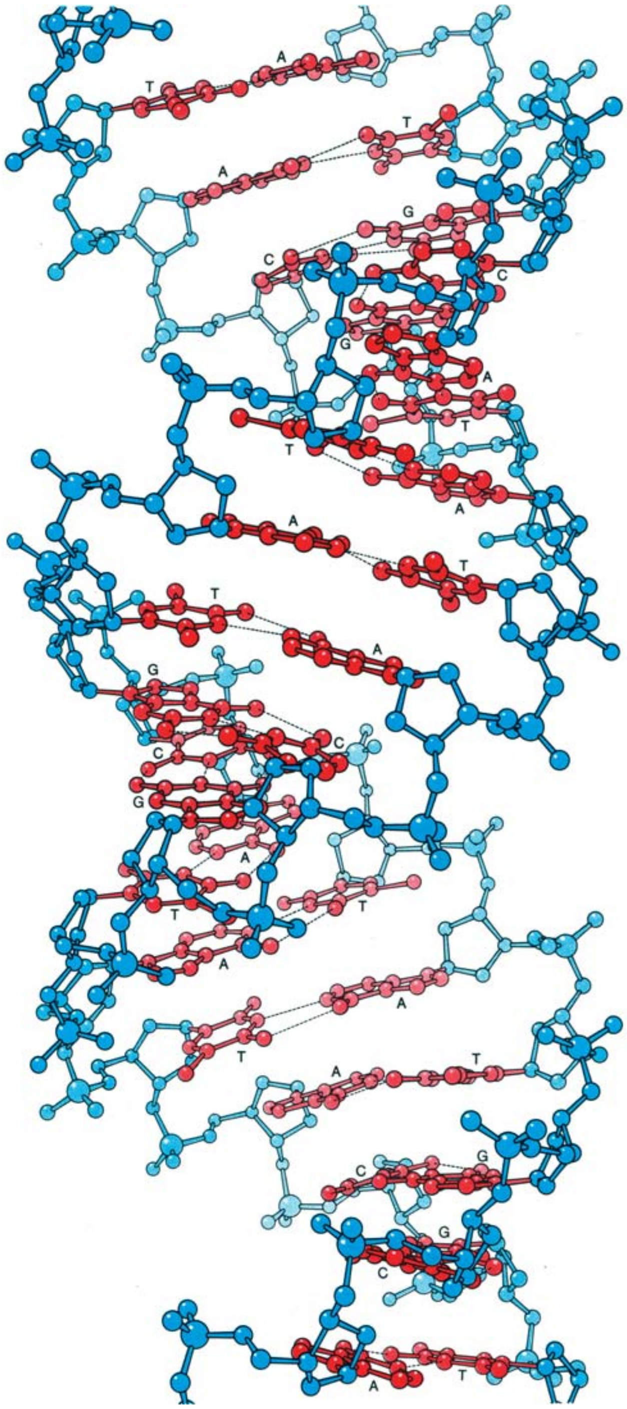


Figure 23.4.1.2

Infinite A-DNA helix, generated from the X-ray crystal structure of the hexamer G-G-T-A-T-A-C-C (references A2 and A7 in Table A23.4.1.1) by deleting the outer base pair from each end and stacking images of the resulting truncated hexamer so their outer phosphate groups overlapped. This generates an endless helix that exhibits the local structural features of the X-ray crystal structure. Note the degree to which the A helix resembles an antiparallel double-stranded ribbon wound around an invisible helical core (the 'hot wire' axis of Fig. 23.4.1.1). (From Dickerson, 1983.) Reprinted courtesy of the estate of Irving Geis. Rights owned by Howard Hughes Medical Institute.

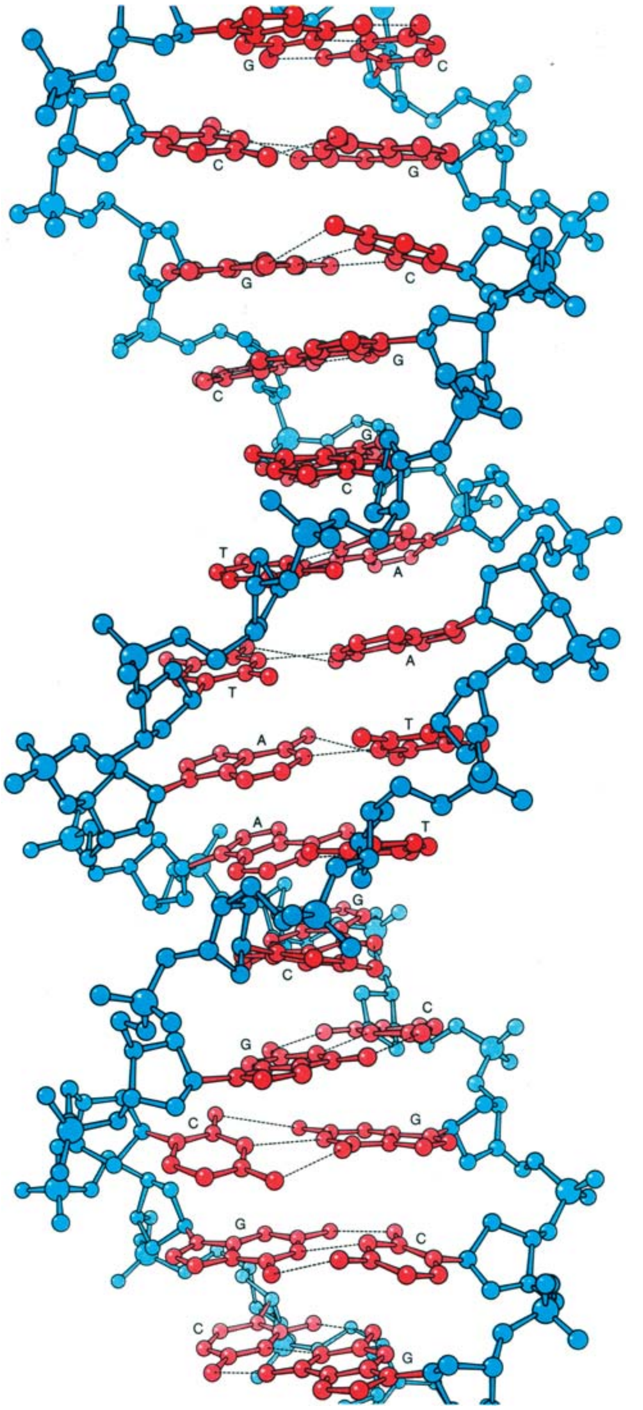


Figure 23.4.1.3

Infinite B-DNA helix, generated in a similar manner to Fig. 23.4.1.2 from the central ten base pairs of the dodecamer C-G-C-G-A-A-T-T-C-G-C-G (B1–B5). Note that the minor groove is narrow in the AT region facing the viewer at the centre, but appreciably wider in the GC regions on the back side of the helix at top and bottom. Propeller twisting, or deviations of bases from coplanarity within one pair, is one sequence-dependent aspect of DNA that was not suspected from the averaged structures obtained from fibres. (From Dickerson, 1983.) Reprinted courtesy of the estate of Irving Geis. Rights owned by Howard Hughes Medical Institute.

T, and G with C. Hence, genetic information present in one strand is passed on to the complementary strand. The standard two-base pairs are shown in Fig. 23.4.2.7 along with the conventional numbering of the atoms. Backbone sugar and phosphate atoms are primed while base atoms are unprimed, as, for example, C1′ and N9 at opposite ends of a purine glycosydic bond. The G·C base pair is held together by three hydrogen bonds, whereas an A·T pair has only two. This means that A·T pairs show less resistance to propeller twisting (counter-rotation of the two bases about their common long axis), and this will have an effect on minor groove width, as seen later. The patterns of hydrogen-bond acceptors (*A*) and donors (*D*) on the major and minor groove edges of base pairs are important elements in recognition of base sequence by drugs and control proteins.

Other related but nonstandard base pairs are compared in Fig. 23.4.2.8. Inosine (I) is useful in studying properties of DNA in

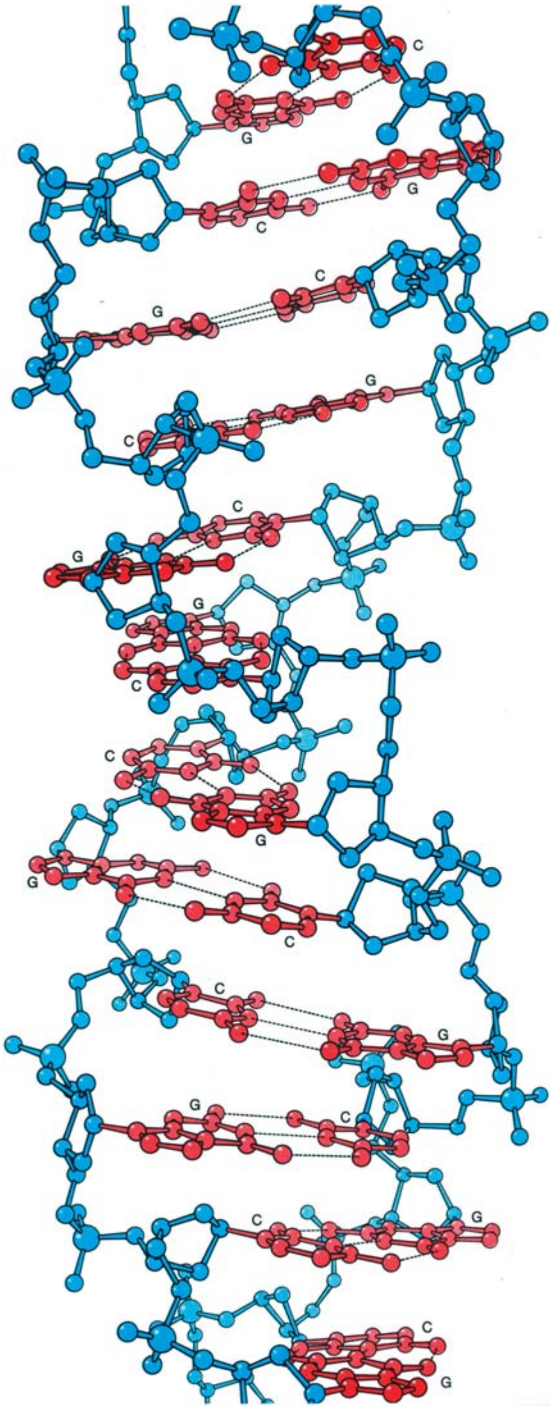


Figure 23.4.1.4
Infinite Z-DNA helix, generated as before from the central four base pairs of the hexamer C-G-C-G-C-G (Z1). G and C bases alternate along each chain. The sugar–phosphate backbone adopts a pronounced zigzag pathway, rising vertically past each guanine, but travelling horizontally across the helix at cytosines. Hence, the formal helix repeat is two base pairs, G followed by C, rather than a single base pair, as in the A and B helices. Note that the structures of Z-DNA and A-DNA are in many ways the inverse of one another. The Z helix is left-handed, tall and slim, with a deep minor groove, a flattened major groove and small propeller twist. The A helix is right handed, short and broad, with a deep major groove, a shallow minor groove and large propeller twist. (From Dickerson, 1983.) Reprinted courtesy of the estate of Irving Geis. Rights owned by Howard Hughes Medical Institute.

that, when paired with cytosine (C), it creates a G·C-family base pair having overall similarity to A·T. Similarly, diaminopurine (DAP) [also known as 2-aminoadenine (2aA)], when paired with thymine (T), creates a G·C-like pair from A·T-family bases.

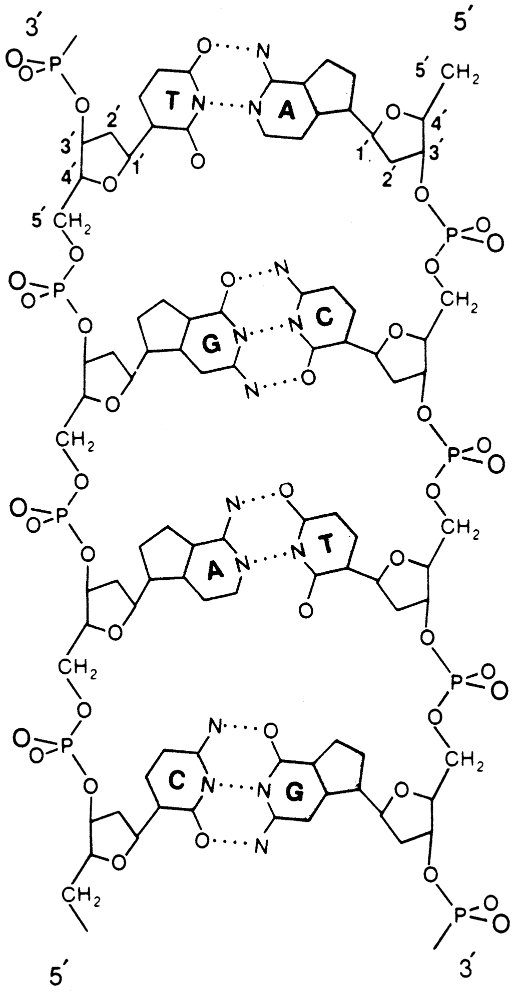


Figure 23.4.2.1
Unrolled schematic of A- or B-DNA, viewed into the minor groove. Paired bases are attached to backbone chains that run in opposite directions: downward on the right and upward on the left. Z-DNA differs from A- and B-DNA in that the two backbone chains run in opposite directions from those shown here. Hence, Z-DNA cannot be obtained from A- or B-DNA by simple twisting around the helix axis.

Hence, in a given experimental situation, one can unscramble the relative significance of number of hydrogen bonds *versus* identity and location of exocyclic groups.

The conventional Watson–Crick base pairing of Fig. 23.4.2.7 uses the hexamer 'end' of the purine base. A different type of base pairing was proposed many years ago by Hoogsteen (1963), in which the upper edge of the purine was used: N7 and N6/O6. Hoogsteen base pairing is shown between the left-hand two bases in each part of Fig. 23.4.2.9. Note that in Hoogsteen base pairing of A and T, each ring provides both a hydrogen-bond donor and an acceptor. Guanine cannot do this, since both its N7 and O6 positions are acceptors. As a consequence, in a G·C pair, C must supply both of the hydrogen-bond donors. It can only form a Hoogsteen base pair with G when the cytosine ring is protonated. This would lead one to expect triplex formation only at low pH. However, the stability of a triplex can, to a certain extent, alter the pK_a of the N—H proton itself. (Recall the shift in pK_a of buried Asp and His groups in the active sites of enzymes.) Hence, with a single-chain DNA, G-A-G-A-G-A-A-C-C-C-C-T-T-C-T-C-T-C-T-T-T-C-T-C-T-C-T-T, that folds back upon itself twice to build a triplex, NMR experiments indicate a significant amount of triplex remaining even at pH 8.0 (Sklenár & Feigon, 1990; Feigon, 1996).

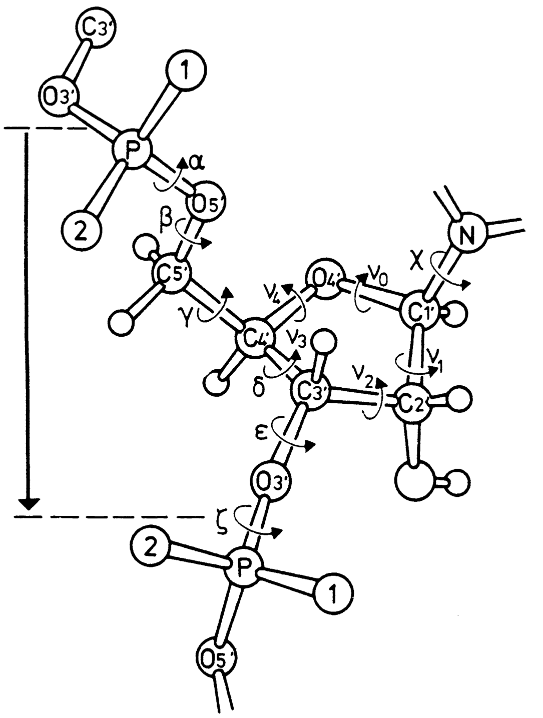


Figure 23.4.2.2
Sugar–phosphate backbone of RNA and DNA polynucleotides. One nucleotide begins at a phosphorus atom and extends just short of the phosphorus atom of the following nucleotide, with the conventional positive direction being $\mathrm{P}\rightarrow\mathrm{O5'}-\mathrm{C5'}-\mathrm{C4'}-\mathrm{C3'}-\mathrm{O3'}\rightarrow\mathrm{P}$, as indicated by the arrows. Main-chain torsion angles are designated α through ζ, and torsion angles about the five bonds of the ribose or deoxyribose ring are ν_0 through ν_4, as shown. If one imagines atoms $\mathrm{O3'}-\mathrm{P}-\mathrm{O5'}$ as a hump-backed bridge, as one crosses the bridge in a positive chain direction, oxygen atom O1 is to the left and O2 is to the right. These oxygens, accordingly, are sometimes designated O_L and O_R. The $-\mathrm{OH}$ group attached to the C2′ atom of the ribose ring in RNA shown here is replaced by $-\mathrm{H}$ in the deoxyribose ring of DNA. Atom N to the right is part of the base attached to the sugar ring: N1 in pyrimidines and N9 in purines. Torsion angle χ is defined by $\mathrm{O4'}-\mathrm{C1'}-\mathrm{N1}-\mathrm{C2}$ in pyrimidines and $\mathrm{O4'}-\mathrm{C1'}-\mathrm{N9}-\mathrm{C4}$ in purines.

23.4.2.4. Helix parameters

An important advantage of single-crystal oligonucleotide structures over fibre-based models is that one can actually observe local sequence-based departures from ideal helix geometry. B-DNA fibre models indicated a mean twist of *ca* 36° per step, or ten base pairs per turn, whereas A-DNA fibre patterns indicated less winding: *ca* 33° per step or 11 base pairs per turn. Twist, rise per base pair along the helix axis, horizontal displacement of base pairs off that axis, and inclination of base pairs away from perpendicularity to the axis are all intuitively obvious parameters. But when single-crystal structures began appearing in great numbers in the mid-1980s, it became imperative that uniform names and definitions be used for these and for less obvious, but increasingly significant, local helix parameters.

An EMBO workshop on DNA curvature and bending, held at Churchill College, Cambridge, in September 1988, led to an agreement on definitions and conventions that was published simultaneously in four journals (Dickerson *et al.*, 1989). Fig. 23.4.2.10 shows the reference frames for two successive base pairs, and Figs. 23.4.2.11 and 23.4.2.12 illustrate local helix parameters involving rotation and translation, respectively. Subsequent experience has shown the most useful parameters to be inclination, propeller, twist and roll among the rotations, and x displacement, rise and slide among the translations. As mentioned at the beginning of this chapter, inclination and x displacement are the two properties that best differentiate A- from B-DNA. The four most widely used computer programs for calculation of local helix parameters are *NEWHELIX* by Dickerson (B7, B46), *CURVES* by Lavery & Sklenar (1988, 1989), *BABCOCK* by Babcock & Olson (Babcock *et al.*, 1993, 1994; Babcock & Olson, 1994) and *FREEHELIX* (Dickerson, 1998*c*). *NEWHELIX* was the earliest of these, but it performs all calculations relative to a best overall helix axis. This is satisfactory for single-crystal DNA structures, but makes the program unusable for the 180° bending observed in some protein–DNA complexes. *CURVES* is especially convenient for mapping the axis of a bent or curved helix. *FREEHELIX*, which evolved from *NEWHELIX*, calculates all parameters relative to local base-pair geometry, without assuming an overall axis, and permits display

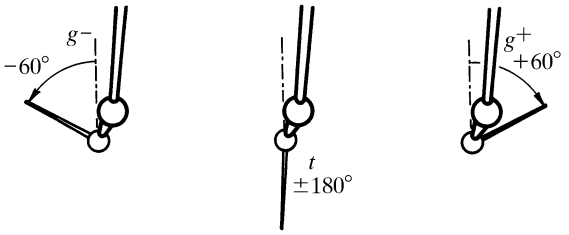


Figure 23.4.2.3
Definition of torsion angles. A positive angle results from clockwise rotation of the farther bond, holding the nearer bond fixed. Torsion angle +60° is designated as *gauche*$^+$ or g^+, angle 180° is *trans* or *t* and angle −60° is *gauche*$^-$ or g^-.

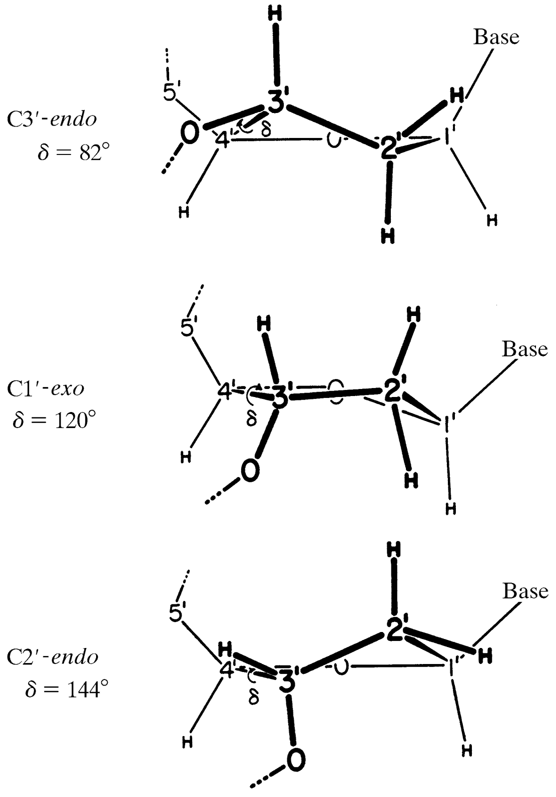


Figure 23.4.2.4
The three most common furanose ring geometries. The planar form of the five-membered ribose or deoxyribose ring is unstable because of steric hindrance from side groups; one of the five atoms prefers to pucker out-of-plane on one side of the ring or the other. Puckering toward the same side of the ring as the C5′ atom is termed *endo*, and puckering toward the opposite 'outside' surface is termed *exo*. The main-chain torsion angle δ is related to sugar ring conformation because of the motion undergone by the $\mathrm{C3'}-\mathrm{O3'}$ bond during changes in puckering.

Table 23.4.2.1
Average torsion-angle properties of A-, B- and Z-DNA (°)

Values listed are mean torsion angles, with standard deviations in parentheses. Conformations are only approximate; — indicates a non-*gauche/trans* conformation. B_{II} and Z_{II} are less common variants. For δ, the sugar ring geometry is quoted in place of *gauche/trans*. χ for B-DNA combines pyrimidines and purines. Values were obtained from a sample of 30 A-DNAs, 34 B-DNAs, 22 Z-DNAs and ten nonstandard DNAs in the Nucleic Acid Database. From Schneider *et al.* (1997).

	α	β	γ	δ	ε	ζ	χ
A-DNA	293 (17)	174 (14)	56 (14)	81 (7)	203 (12)	289 (12)	199 (8)
Conformation	g^-	t	g^+	C3′-*endo*	t	g^-	t
B-DNA	298 (15)	176 (9)	48 (11)	128 (13)	184 (11)	265 (10)	249 (16)
Conformation	g^-	t	g^+	C1′-*exo*	t	g^-	g^-
B_{II}-DNA		146 (8)		144 (7)	246 (15)	174 (14)	271 (8)
Conformation		—		C2′-*endo*	g^-	t	g^-
Z_I-DNA – purines	71 (13)	183 (9)	179 (9)	95 (8)	95 (8)	301 (16)	63 (5)
Conformation	g^+	t	t	O4′-*endo*	g^+	g^-	g^+
Z_{II}-DNA – purines					189 (12)	52 (14)	58 (5)
Conformation					t	g^+	g^+
Z_I-DNA – pyrimidines	201 (20)	225 (16)	54 (13)	141 (8)	267 (9)	75 (9)	204 (98)
Conformation	t	—	g^+	C2′-*endo*	g^-	g^+	t
Z_{II}-DNA – pyrimidines	168 (16)	166 (14)					
Conformation	t	t					

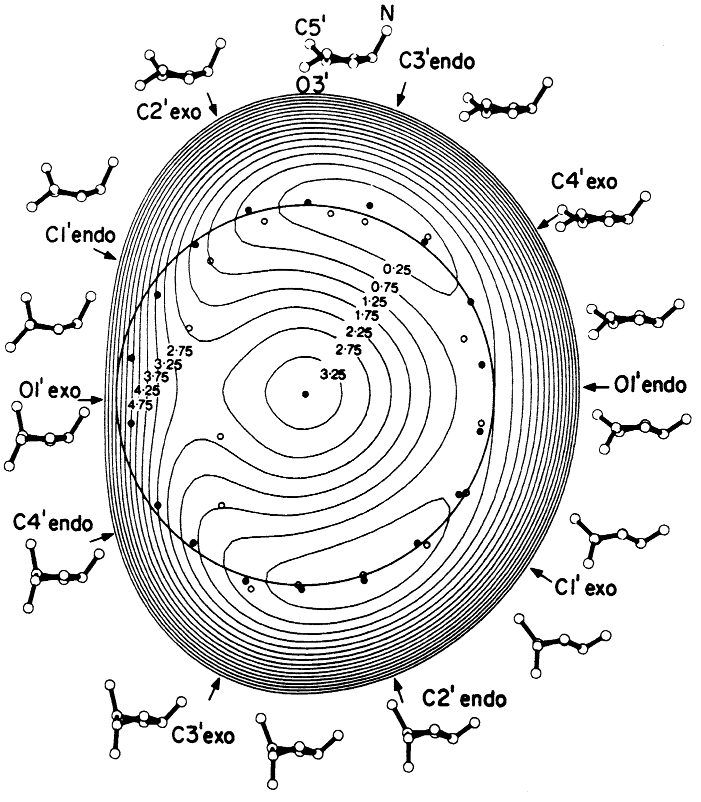


Figure 23.4.2.5
Potential plot of all furanose ring conformations. Energies are in kcal mol^{-1}. The distance from the central point gives the maximum displacement of the out-of-plane atom from the plane of the other four. The circle is a constant-displacement trajectory chosen to pass through the potential minima on the right three-quarters of the plot. C2′-*endo* and C3′-*endo* are especially favoured, whereas O1′-*exo* on the left is highly disfavoured. The path from C2′-*endo* through C1′-*exo*, O1′-*endo* and C4′-*exo* to C3′-*endo* is a low-energy path, and many examples all along this path are known in B-DNA helices. Reprinted with permission from Levitt & Warshel (1978). Copyright (1978) American Chemical Society.

Table 23.4.2.2
Sugar ring conformations, pseudorotation angles and torsion angle δ

Ring conformation	Pseudorotation angle (°)	Torsion angle δ (°)
C3′-*endo*	18	82
C4′-*exo*	54	82
O4′-*endo*	90	96
C1′-*exo*	126	120
C2′-*endo*	162	144
C3′-*exo*	198	158
C4′-*endo*	234	158
O4′-*exo*	270	144
C1′-*endo*	306	120
C2′-*exo*	342	96

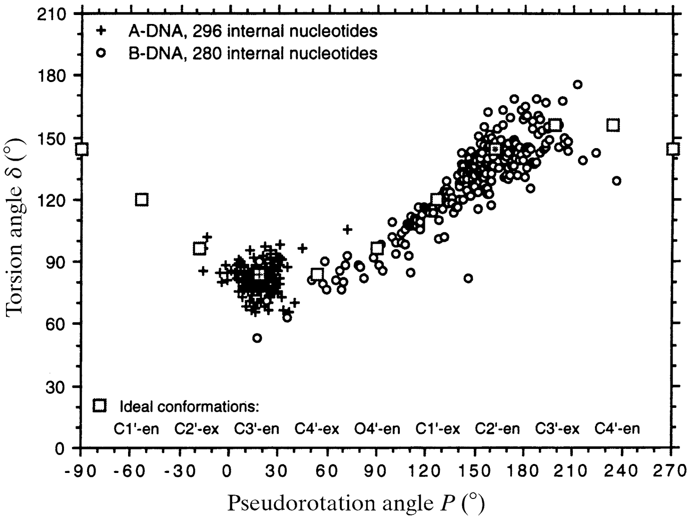


Figure 23.4.2.6
Plot of observed sugar conformations in 296 nucleotides of A-DNA (crosses) and 280 of B-DNA (open circles). Open squares mark ideal relationships between torsion angle δ (vertical axis) and pseudorotation angle P (horizontal axis) from the expression $\delta = 40° \cos(P + 144°) + 120°$. Deviations from this ideal curve for real helices arise, because the amplitude of pseudorotation (or displacement of one atom from the mean plane of the others) varies from one ring to another. Note the tight clustering of A-DNA points around C3′-*endo* and the broader distribution of B-DNA conformations.

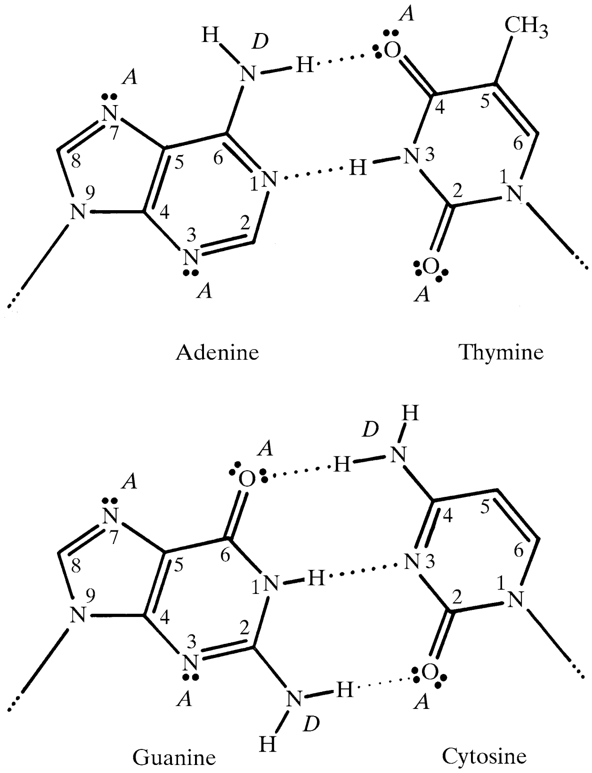


Figure 23.4.2.7
A·T and G·C base pairs with minor groove edge below and major groove edge above. *A* is a hydrogen-bond acceptor, *D* is as hydrogen-bond donor.

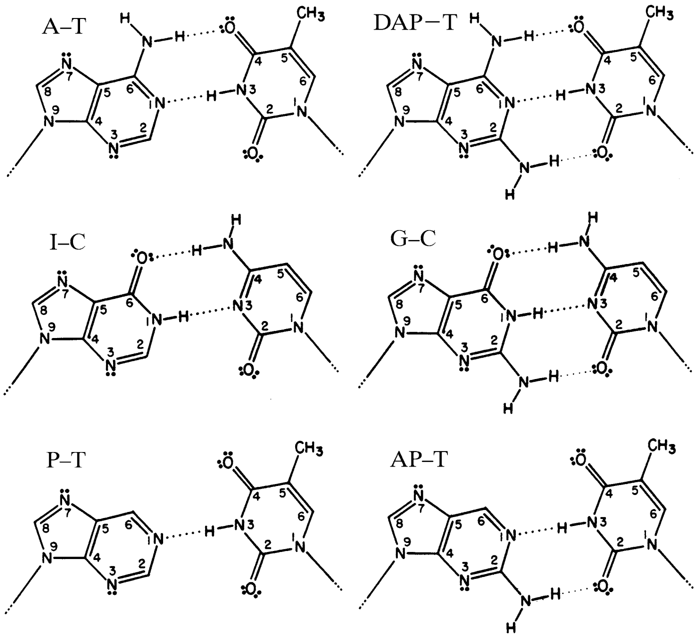


Figure 23.4.2.8
Alternative purines and pyrimidines, and possible base pairings. Purines: P = purine; AP = 2-aminopurine; A = adenine or 6-aminopurine; DAP = 2,6-diaminopurine (also known as 2aA = 2-aminoadenine); G = guanine; I = inosine. Pyrimidines: T = thymine (uracil if methyl group is absent); C = cytosine. DAP–T is a nonstandard AT-family analogue of G–C, and I–C is a nonstandard GC-family analogue of A–T.

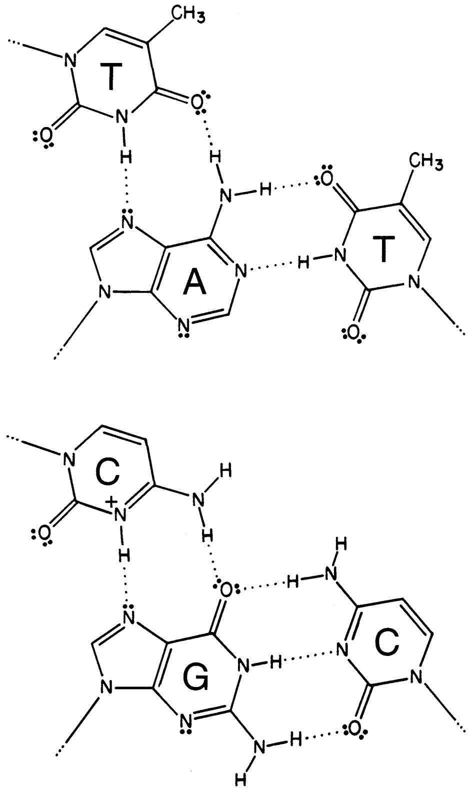


Figure 23.4.2.9
Watson–Crick pairing of a purine (A or G) with a pyrimidine to its right (T or C), and Hoogsteen pairing of the same purine with a pyrimidine above it. This combination of Watson–Crick and Hoogsteen pairing is found in triple helices or triplexes. Note that Hoogsteen pairing of G and C can only occur at a pH at which C is protonated, because the extra proton is essential for the second hydrogen bond.

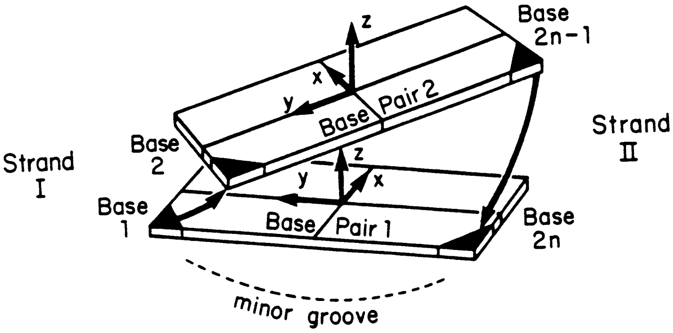


Figure 23.4.2.10
Definitions of local reference axes (x, y, z) at the first two base pairs of an n-base-pair double helix. Base 1 is paired with base $2n$, base 2 with base $2n - 1$ *etc.* Shaded corners represent attachment points to sugar rings. Curved arrows denote 5′-to-3′ 'positive' directions of each backbone chain. Note that when looking into the minor groove, as here, the two strands illustrate a clockwise rotation, upwards on the left and downwards on the right. This is true for A- and B-DNA, but for Z-DNA, the sense of the two backbone strands is reversed.

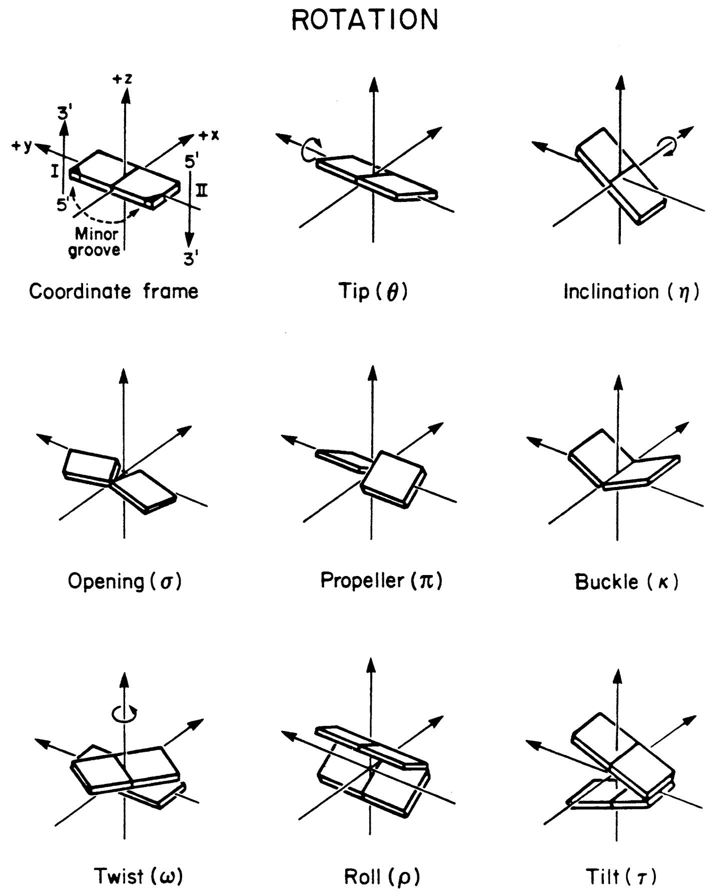

ROTATION
+z
3'
+y
+x
5'
I
II
5'
Minor groove
3'
Coordinate frame
Tip (θ)
Inclination (η)
Opening (σ)
Propeller (π)
Buckle (κ)
Twist (ω)
Roll (ρ)
Tilt (τ)

Figure 23.4.2.11
Local helix parameters involving rotations. Tip and inclination describe the orientation of a base pair relative to the helix axis, produced by rotation about the base-pair long axis or short axis, respectively. Opening, propeller and buckle describe rotations of the two bases of a pair relative to one another. Twist, roll and tilt describe changes of orientation from one base pair to the next, *via* rotations about the *z*, *y* and *x* axes, respectively.

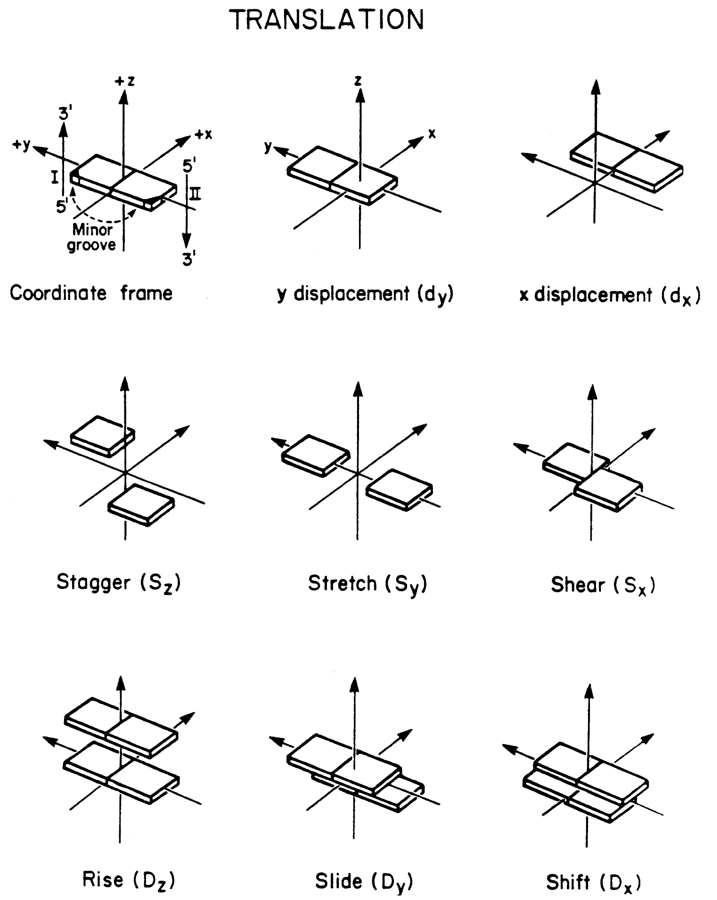

TRANSLATION
+z
3'
+y
+x
5'
I
II
5'
Minor groove
3'
z
y
x
Coordinate frame
y displacement (dy)
x displacement (dx)
Stagger (Sz)
Stretch (Sy)
Shear (Sx)
Rise (Dz)
Slide (Dy)
Shift (Dx)

Figure 23.4.2.12
Local helix parameters involving translations. *y* and *x* displacements describe shifts of a lone base pair along its long or short axis, respectively. Stagger, stretch and shear describe displacements of the two bases of a pair relative to one another. Rise, slide and shift describe displacements from one base pair to the next, *via* translations along the *z*, *y* and *x* axes, respectively.

of normal vector plots that are especially useful in analysing bending in DNA–protein complexes (Dickerson & Chiu, 1997).

23.4.2.5. Syn/anti glycosyl bond geometry

The glycosyl bond angle, χ, about the bond connecting a sugar ring to a base is a special case of torsion angle, and is defined by O4′—C1′—N1—C2 for pyrimidines and O4′—C1′—N9—C4 for purines. In A- and B-DNA, the normal range of χ is 160 to 300°. This is known as the *anti* conformation (right-hand side of Fig. 23.4.2.13) and swings the sugar ring out away from the minor groove edge of the base pair. In Z-DNA, pyrimidines also exhibit the *anti* glycosyl bond conformation, but purines adopt the *syn* geometry shown on the left-hand side of Fig. 23.4.2.13. Now the sugar ring is rotated so that it intrudes into the minor groove, and χ lies in the range 50 to 90°.

23.4.3. Comparison of A, B and Z helices

Figs. 23.4.3.1–23.4.3.3 show the original stereo pairs that were redrawn by Irving Geis in preparing Figs. 23.4.1.2–23.4.1.4. These stereo pairs were constructed from X-ray structures of A-, B- and Z-DNA oligomers by deleting the outermost base pair from each end, eliminating the backbone as far as the first phosphate group, and then stacking these trimmed-down helices on top of one another, with phosphate groups overlapping, to create an infinite

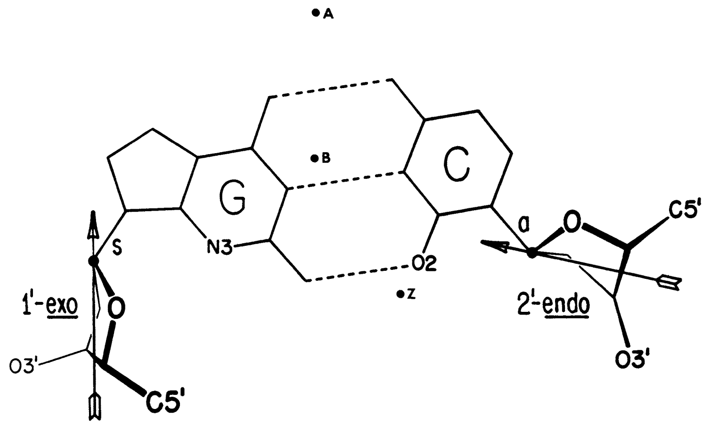

•A
•B
•Z
G
C
N3
O2
s
a
O
O
C5'
C5'
1'-exo
2'-endo
O3'
O3'

Figure 23.4.2.13
Syn versus anti orientation about the glycosyl bond connecting sugar and base. Right: *anti* conformation, with χ *ca* 210°. Left: *syn* conformation, with χ around 60°. Both A- and B-DNA only employ the *anti* geometry; Z-DNA uses *anti* for pyrimidines and *syn* for purines, as shown here. Note that the 5′-to-3′ direction in both rings is down into the paper. Hence, antiparallel backbone chains can be achieved only by a zigzag chain geometry with local chain reversals, as shown later in Fig. 23.4.3.4. Black dots labelled A, B and Z indicate the position of the helix axis relative to the base pairs in A-, B- and Z-DNA.

helix. They are improvements over the idealized infinite helices generated from fibre diffraction in that they display local variation in helix parameters that only single-crystal analyses can reveal. In the present context, they are good subjects for discussion of the differences between the three helix types.

23.4.3.1. *x* displacement and groove depth

A-DNA (Wahl & Sundaralingam, 1996, 1998), B-DNA (Berman, 1996; Dickerson, 1998*b*) and Z-DNA (Ho & Mooers, 1996; Basham *et al.*, 1998) have each been the subject of recent reviews, to which the reader is referred for details that cannot be covered here. The distinctive properties of the three helices are listed in Table 23.4.3.1. The most obvious distinction is handedness: A and B are right-handed helices, whereas Z is left-handed. Moreover, the position of each base pair relative to the helix axis is quite different. As noted in Fig. 23.4.2.13, the helix axis passes through base pairs in B-DNA, lies on the minor groove side of base pairs in Z-DNA, and on the major groove side in A-DNA. In terms of the helix parameters of Fig. 23.4.2.12, A-DNA has a typical *x* displacement of d_x = +3 to +5 Å, B-DNA has d_x = −1 to 0 Å, and Z-DNA has d_x = −3 to −4 Å. There is virtually no overlap between these three ranges; *x* displacement, d_x, in fact, is a better criterion for differentiating the three classes of helix than is sugar ring conformation.

A direct consequence of these *x* displacement values is great differences in depths of major and minor grooves. Both grooves are of equivalent depth in B-DNA because base pairs sit on the helix axis. In A-DNA, a base pair is pushed off-axis so that its minor edge approaches the helix surface, making the minor groove very shallow and the major groove cavernously deep. In Z-DNA, it is the major edge of each base pair that is pushed toward the surface, so that the minor groove is deep and the major groove is so shallow as hardly to be characterized as a groove at all. It is sometimes stated that 'Z-DNA has no major groove', but space-filling stereos, such as Fig. 1 of reference Z6 or Fig. 3 of Z23 reveal the shallowest of major grooves running around the helix cylinder, flanked by very slightly higher phosphate backbones.

23.4.3.2. Glycosyl bond geometry

In both A- and B-DNA, all glycosydic bonds are *anti*, with sugar rings swung to either side away from the minor groove, as in Fig. 23.4.3.4(*a*). As mentioned earlier, when viewed into the minor groove, the backbone chains describe a clockwise rotation, with the chain on the right running downward, and that on the left upward, as in Fig. 23.4.2.1. In Z-DNA, both chains run in the *opposite* direction, leading to a counterclockwise rotation sense viewed into the minor groove. But Z-DNA has yet another striking (and defining) feature. Purines and pyrimidines alternate along each chain. G and C are most strongly favoured by far, but A and T can substitute intermittently at a price in stability. Breaking the strict alternation of purines and pyrimidines is even more unfavourable and is rarely encountered in crystal structures (Table A23.4.1.3). At each purine base, the glycosyl bond is rotated into the minor groove to the *syn* position, as in Fig. 23.4.3.4(*c*). This causes the local backbone

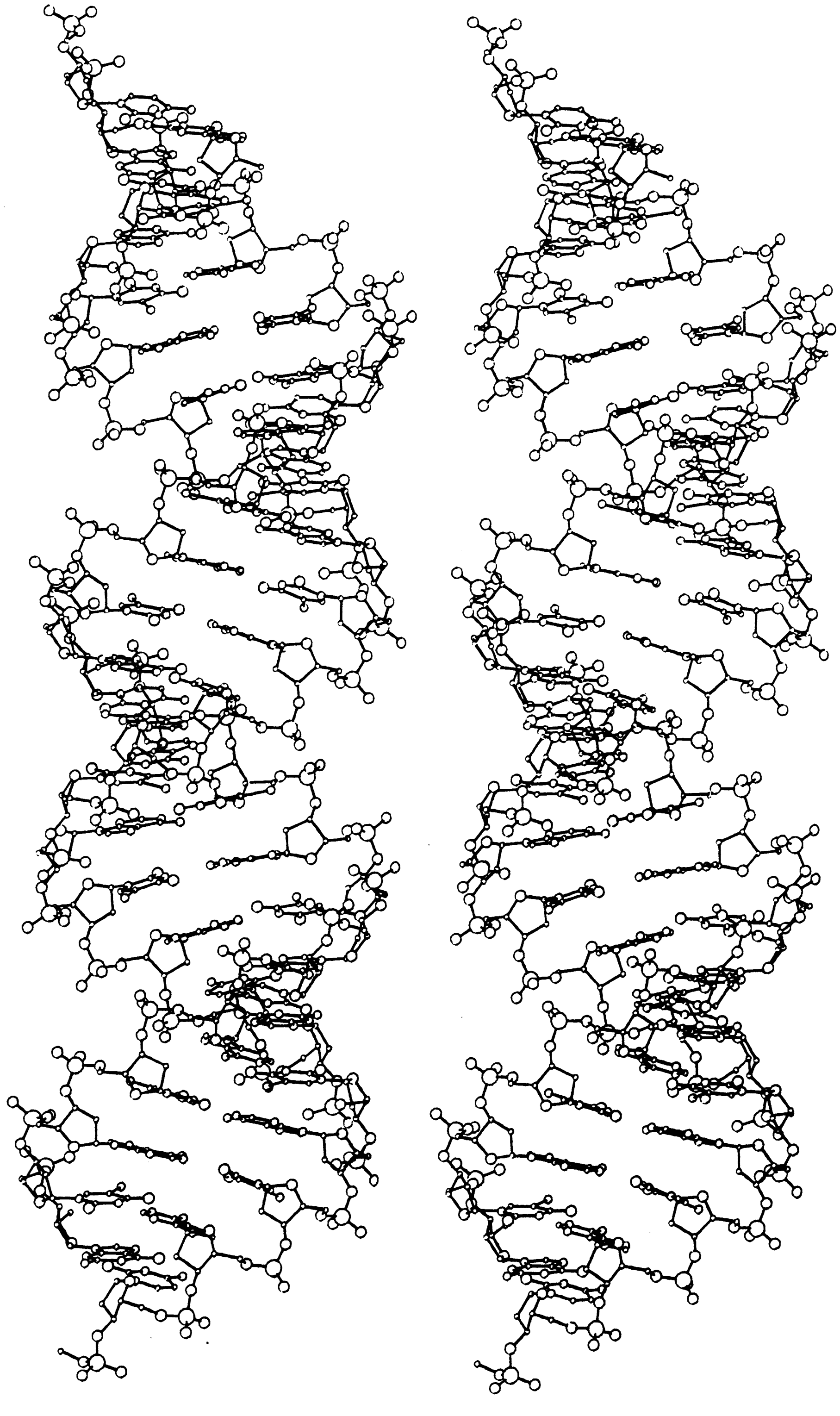


Figure 23.4.3.1
The A-DNA stereo pair drawing from which Fig. 23.4.1.2 was derived, with repeating sequence -(G-T-A-T-A-C)$_n$-. The impression of the A helix as a ribbon wrapped around an imaginary core is even more strongly developed in this stereo. (From Dickerson, 1983.)

directions, defined by sugar ring atoms C4′ and C3′, to be parallel in the two strands. Z-DNA avoids becoming a parallel-chain helix by performing a local chain reversal at each pyrimidine. In Fig. 23.4.3.4(*c*), although the local C4′–C3′ chain direction at the cytosine sugar is downward, the double loop in backbone chain gives it a net upward orientation. In stereo Fig. 23.4.3.3, the ascending backbone chain rises smoothly past each guanine, with a chain path parallel to the helix axis. However, the chain bends abruptly at right angles when passing a cytosine, in a direction tangential to the helix cylinder. Guanine sugar rings point their O4′ oxygen atoms in the backward chain direction (as is also true for all bases in A- and B-DNA), but cytosine sugars point their oxygens in the forward direction. This 'up at G, across at C' pathway and inversion of sugar rings is what produces the zigzag backbone pathway that leads to the name Z-DNA. The O4′ atom of each cytosine sugar is stacked on top of the guanine ring of the subsequent nucleotide, and this stacking of a polar O (or N) on top of a polarizable aromatic ring contributes to the stability of the Z helix, as it does to many other base–base interactions to be discussed later (Bugg *et al.*, 1971; Thomas *et al.*, 1982; B32).

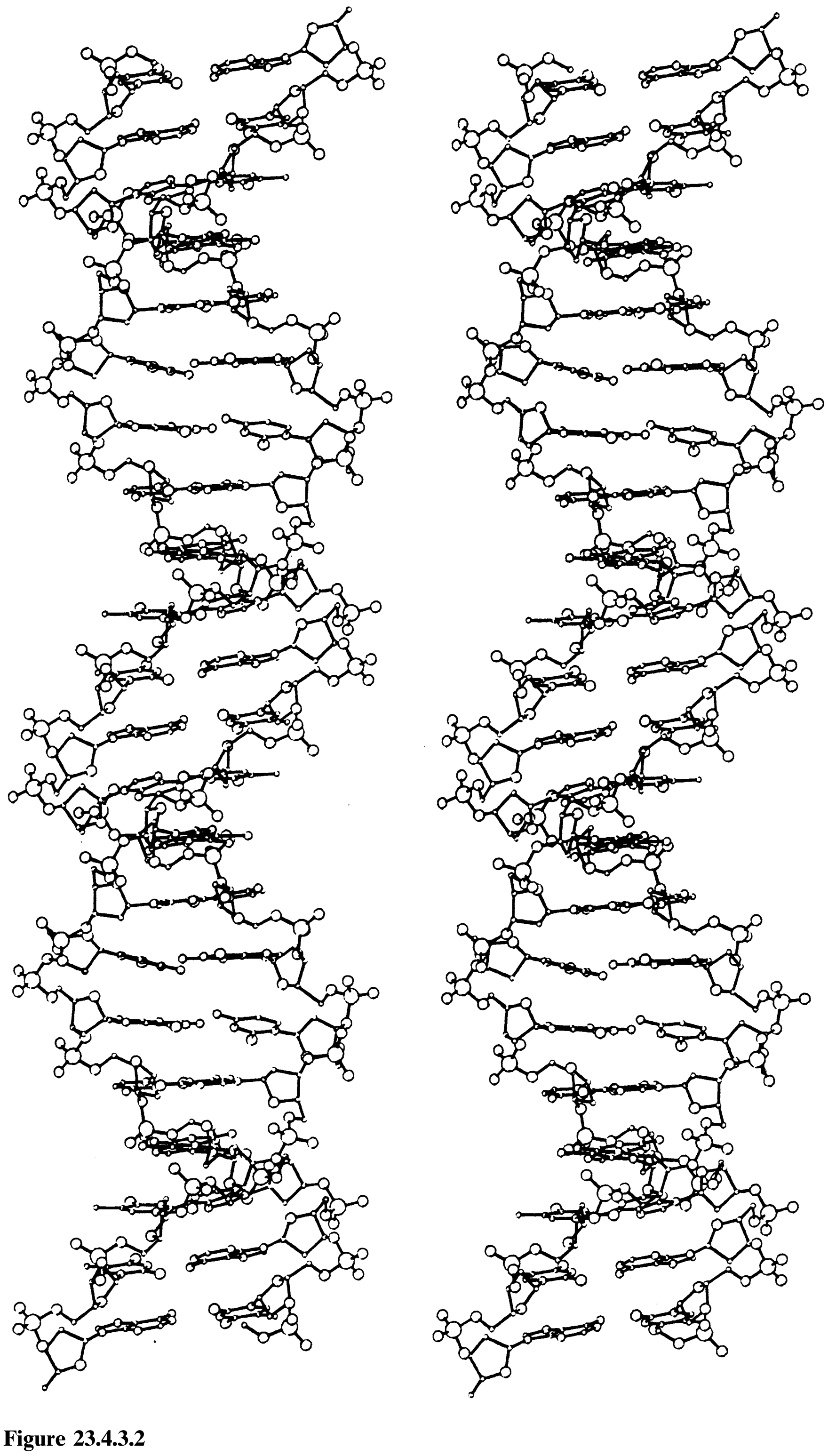


Figure 23.4.3.2
The B-DNA stereo pair drawing from which Fig. 23.4.1.3 was derived, with repeating sequence -(G-C-G-A-A-T-T-C-G-C)$_n$-. The variation of minor groove widths on the front and back sides of the helix is striking. (From Dickerson, 1983.)

23.4.3.3. *Sugar ring conformations*

Sugar ring conformations in A- and B-DNA have a logical structural basis. The B-DNA backbone is more extended than the A-DNA backbone, with P–P distances of *ca* 6.6 Å along one chain, compared with *ca* 5.5 Å in A-DNA. In turn, C2′-*endo* is a more extended ring conformation than C3′-endo, demonstrable in Fig. 23.4.2.4 by a greater distance between C5′ and O3′ atoms. Hence, it is logical that the more extended ring conformation should be associated with the more extended backbone chain. In Z-DNA, the extended C2′-*endo* form is adopted at cytosine, where a zigzag double chain reversal must be accommodated, while the more compact C3′-*endo* occurs at the straight backbone segment running past a guanine.

The cramped *syn* glycosyl conformation is strongly disfavoured, although not absolutely forbidden, at pyrimidines, most probably because of steric clash between the pyrimidine O2 and the *syn* ring (Haschmeyer & Rich, 1967; Davies, 1978; Ho & Mooers, 1996; Basham *et al.*, 1998). Hence, the Z-DNA helix is effectively limited to alternating pyrimidine/purine sequences, with a price that must be paid for intermittent substitution of A and T for G and C, and an even higher price paid for breaking the pyrimidine/purine alternation. This is reflected in the X-ray crystal structures listed in Table A23.4.1.3. Only one non-alternating sequence has been completely solved and published: *C-G-G-G-*C-G (Z40), where adoption of the Z form has been forced by 5-methylation of cytosines (*C). A second non-alternating sequence that includes AT base pairs, *C-G-A-T-*C-G (Z13), was solved in 1985, but its coordinates have never been made public. It, too, required methylation of cytosines to induce the Z form. A third sequence, C-C-G-C-G-G (Z42), opens its terminal base pairs to make intermolecular base pairs with crystal neighbours. The 52 remaining Z-DNA structures in Table A23.4.1.3 all have strict alternation of pyrimidines and purines.

23.4.3.4. *Helical twist and rise, and propeller twist*

The helical repeat unit in Z-DNA is therefore two successive base pairs, rather than the single base pair of A- and B-DNA. Ho & Mooers (1996) propose that the C-G or $^{5'}$pyrimidine-P-purine$^{3'}$ step be considered the fundamental unit of the Z-helical structure, because of the tight overlap between the two base pairs. As can be seen in Fig. 23.4.3.3, in a C-G step the pyrimidine rings from the two base pairs actually stack over one another, whereas the purine rings are packed against neighbouring sugar O4′ atoms. Helix-axis rotation at this step is only −8°, whereas the preceding and following G-C steps have a mammoth −52° twist. Hence, although Z-DNA has 12 base pairs per turn, it technically is not a dodecamer helix, but a hexamer with a two-base-pair repeating unit and a total rotation of −60° per unit.

This virtual restriction to purine/pyrimidine alternation means that Z-DNA cannot be involved in the coding of genetic information. A and B helices have no such restriction; their structures can accommodate a random sequence of bases. Average twist angles are as shown in Table 23.4.3.1, although extreme variation in twist is observed at individual steps in single-crystal structure analyses, from as little as 16° to as much as 55°. Base-sequence preferences for local helix parameters are discussed below.

In both B and Z helices, base pairs are very nearly perpendicular to the helix axis, whereas in the winding double ribbon of A-DNA, the long axis of each base pair is inclined by 10 to 20° away from perpendicularity to the axis. Hence, the rise per base pair for all B-helical steps and for G-C steps of Z-DNA is equal to the thickness of a base pair, 3.4 Å. The rise at a C-G step of Z-DNA is larger because it involves stacking of a sugar oxygen on each purine ring, not ring stacked on ring. For A-DNA, the rise along the helix axis can actually be less than the thickness of a base pair, because adjacent base pairs are stacked at an incline. The perpendicular distance from one base pair to the next in A-DNA is still 3.4 Å. Both A- and B-DNA exhibit considerable base pair propeller twist, especially at A·T pairs with only two hydrogen bonds rather than three. In contrast, Z-DNA, with predominately G·C pairs, shows only a small propeller twist.

The stacking of base pairs has immediate consequences for crystal growth. For Z-DNA, four base pairs are one-third of a helical turn, and six base pairs are a half turn. Hexamers are the most common crystal form in Table A23.4.1.3 by a large majority. In contrast, octamers and decamers are not simple fractions of a turn, and they stack in a disordered manner. One would predict that dodecamers of Z-DNA might crystallize well if the oligomers were not so long as to fall prey to cylindrical disorder.

By the same principles, B-DNA decamers stack easily and well to build pseudo-infinite helices through the crystal, with ordered cylindrical rods packed in six different space groups. The other common crystallization mode for B-DNA, the dodecamer, has a two-base-pair overlap of ends that both stablizes the crystals and yields a functional ten-base-pair repeat. (See Fig. 2 of Dickerson *et al.*, 1987.) Because the dodecamers are held by their outer two base pairs, the central eight pairs are unobstructed and accessible in the crystal, making dodecamers particularly good subjects for the study of minor-groove binding drugs.

A-RNA duplexes [Table A23.4.1.1, part (*k*)] also stack end-for-end in a manner simulating an infinite A helix, even though the end base pairs are inclined and are not perpendicular to the helix axis. This behaviour has been seen for octamers with roughly two-thirds of a helical turn, for nonamers, and for dodecamers with roughly a full turn.

In contrast, crystals of A-DNA behave quite differently. Regardless of chain length, A-DNA helices crystallize with the outer base pair of one helix packed against one wall of the broad, open and relatively hydrophobic minor groove of another helix. This packing mode is sufficiently adaptable to accommodate duplexes of lengths four, six, eight, nine, ten and 12 base pairs. Hence, A-DNA does not

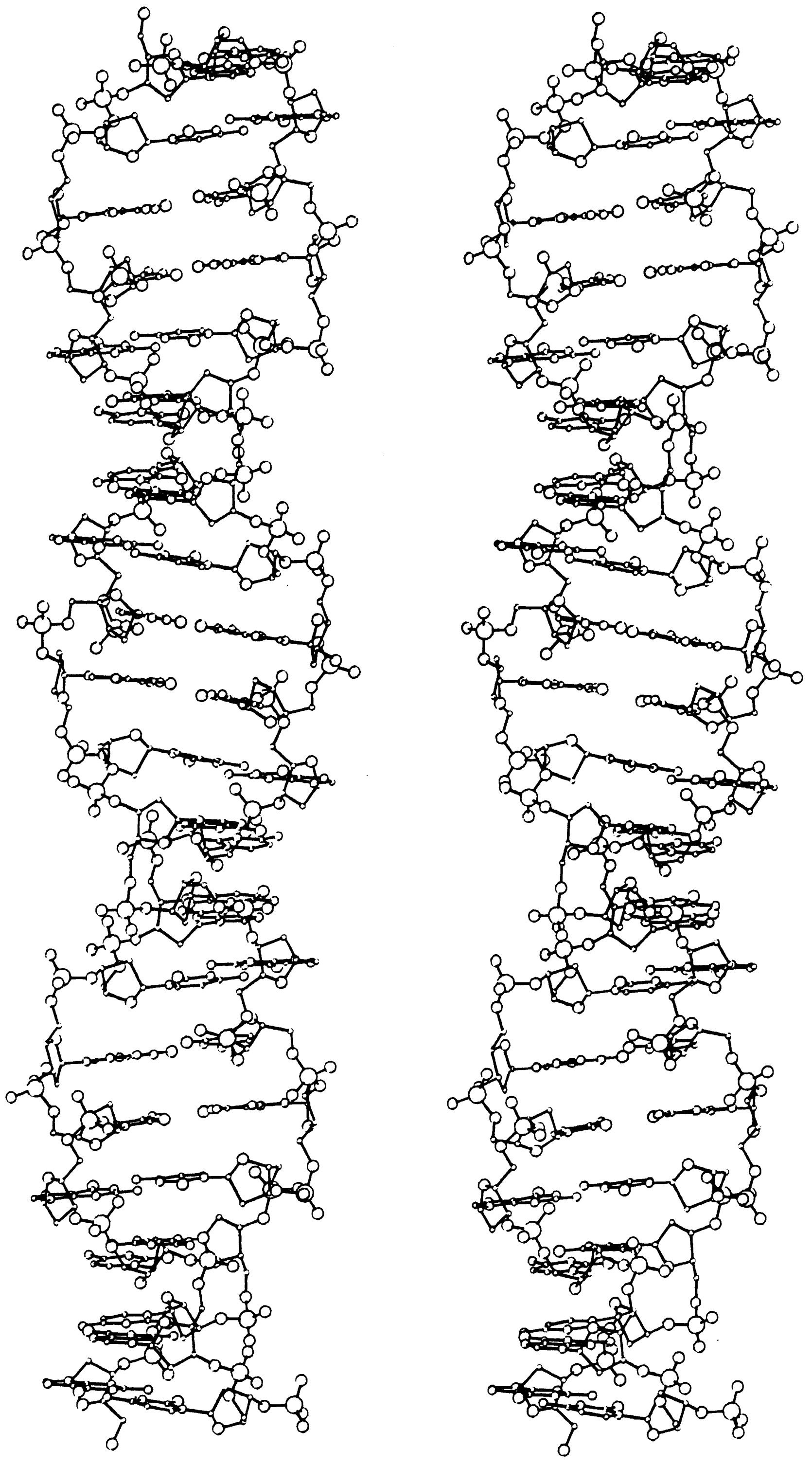


Figure 23.4.3.3
The Z-DNA stereo pair drawing from which Fig. 23.4.1.4 was derived, with repeating sequence -(G-C-G-C)$_n$-. Note the left-handed zigzag path of the sugar–phosphate backbone, which led to its designation as the Z helix. (From Dickerson, 1983.)

Table 23.4.3.1
Comparison of structures of A, B and Z helices

	A	B	Z
Handedness	Right	Right	Left
Helix axis relative to base pairs	Major groove side	Through centre of base pair	Minor groove side
Major groove	Very deep and narrow	Wide, same depth as minor	Very shallow and broad
Minor groove	Shallow and broad	Variable, same depth as major	Very deep and narrow
Glycosydic bonds	*anti*	*anti*	C: *anti* G: *syn*
Minor groove backbone chain sense†	Clockwise	Clockwise	Counterclockwise
Sugar conformation	C3′-*endo* (narrow range)	C1′-*exo*/C2′-*endo* (broad range)	C: C2′-*endo* G: C3′-*endo*
Base pairs per helix repeat	1	1	2
Base sequence limitations	None	None	Alternating (C-G)$_n$ or close variants
Rise per base pair (average)	2.9 Å	3.4 Å	C-G: 4.1 Å G-C: 3.5 Å
Base pair inclination	10–20°	*ca* 0°	*ca* 0°
Mean twist angle	30–33°	34–36°	C-G: −8° G-C: −52°
Helix repeats per turn	11–12	10–10.5	6 (2 base pairs)
Propeller twist	Often substantial, 0–25°	Often substantial, 0–25°	Usually small
Common biological occurrence	RNA	DNA	None?

† Relative 5′-to-3′ directions of the two backbone chains, when viewed into the minor groove.

simulate infinite helices through the crystal lattice, as A-RNA and B- and Z-DNA do.

23.4.3.5. *Allowable RNA helices*

So far this discussion has only been concerned with DNA. Which of the three helix types can be adopted by RNA? Fig. 23.4.3.5 shows that addition of a 2′-OH group to a B-DNA helix [part (*a*)] creates severe steric clash with the phosphate group and sugar ring of the following nucleotide, whereas in an A helix [part (*b*)], the added hydroxyl group extends radially outward from the helix cylinder and causes no steric problems. Hence, the natural helical form for RNA is the A helix, not the B helix. Table A23.4.1.1 shows several single-crystal analyses of A-RNA and RNA/DNA hybrids; Table A23.4.1.2 shows no B-RNA structures. One RNA/DNA hybrid is known as a Z helix: C-G-c-g-C-G (Z24), in which the two central nucleotides are RNA. If one mentally adds an —OH to each C2′ atom in Fig. 23.4.3.3, on the same side of the ring as O3′, it is apparent that the C2′-OH is not inherently incompatible with the Z helix, as it is with the B helix. At guanine sugars, the C2-OH points out and away from the helix, while at cytosine sugars it points away from the base into the spacious minor groove.

23.4.3.6. *Biological applications of A, B and Z helices*

The B helix is the biologically relevant structure for DNA. The A form might logically be adopted at the stage of transient DNA–RNA duplexes during transcription, but elsewhere the B form holds sway. It was once thought that binding of DNA to a protein surface, most particularly nucleosomal winding, might constitute a sufficient dehydration of bound water molecules from the DNA duplex to shift it to the A form. This proved to be false; nucleosomal DNA clearly retains the B conformation. The closest that one comes to biological A-DNA is local deformations upon binding of B-DNA to a few proteins that have been described as ‘A-like distortions’. On the other hand, the A helix has been found repeatedly in RNA duplexes, including tRNA and ribozymes.

The situation is even more restrictive with the Z helix. Although its alternating purine/pyrimidine sequence makes it unusable for genetic *coding*, the suggestion has been made on many occasions that Z-DNA might be an important element in genetic *control* by being involved in negative supercoiling (Herbert & Rich, 1996). It has been shown that a left-handed DNA conformation can be induced by negative superhelical stress, but it is not absolutely clear that this induced, left-handed

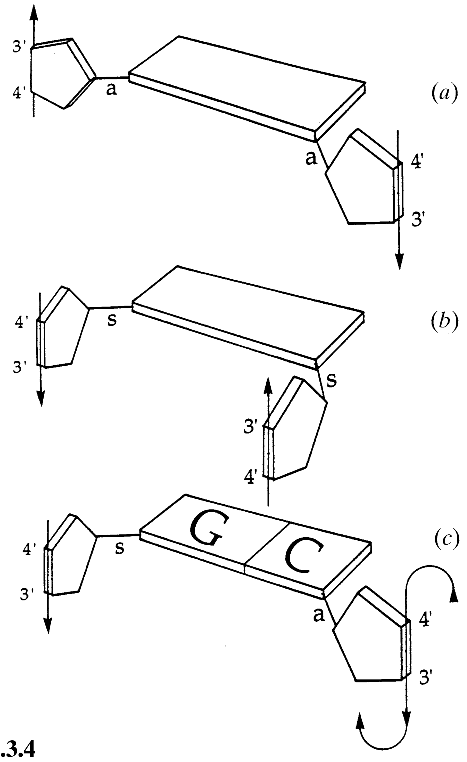


Figure 23.4.3.4
Glycosyl conformation and chain sense. (*a*) Glycosyl conformations *anti*/*anti*, backbone chains antiparallel, with clockwise sense when viewed into the minor groove, as here. This is typical for A- and B-DNA. (*b*) Glycosyl conformation *syn*/*syn*, backbone chains antiparallel, with counterclockwise sense viewed into minor groove. This is not known for any nucleic acid duplex. (*c*) Glycosyl conformation *syn* at G and *anti* at C, with the C4′—C3′ edge of the sugar pointing downward in both strands, which would seem to imply a parallel-stranded helix. However, in Z-DNA, antiparallel strands are achieved by a local reversal of chain direction at each C, as shown here. This produces the zigzag backbone pathway that is characteristic of the Z helix, visible in Fig. 23.4.3.3.

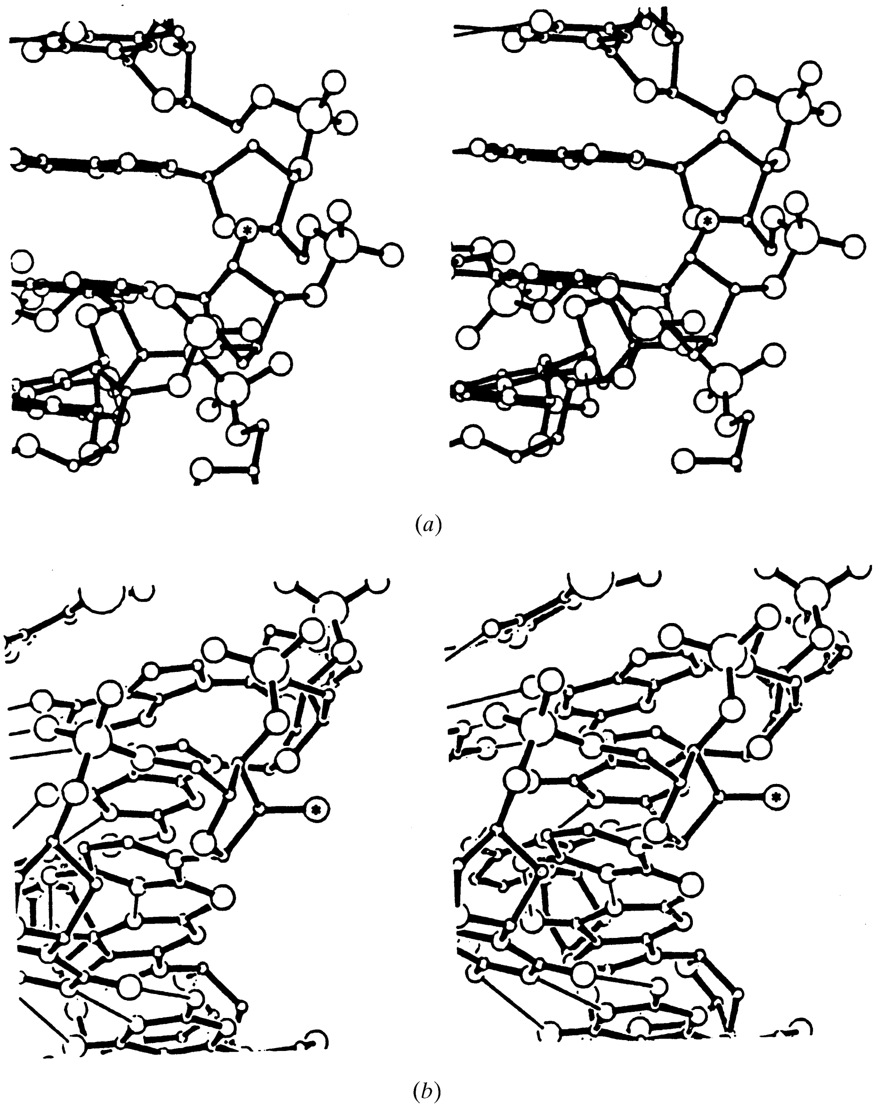


Figure 23.4.3.5
The role of the C2′-OH in RNA helix geometry. (*a*) Addition of a C2′-OH group (*) to the B-DNA helix leads to close contacts and unallowable steric hindrance with the following O5′ and O4′ atoms and, to a lesser degree, the subsequent base itself. (*b*) A C2′-OH group added to the A-DNA helix extends outward radially from the helix cylinder surface and produces no steric clashes. Hence, A-RNA is quite possible, whereas B-RNA is disallowed.

conformation is the same as the Z helix seen in crystal structures of small oligomers. As noted by Herbert & Rich (1996), after nearly twenty years of enquiry, it is still far from certain that Z-DNA itself has any demonstrable biological role.[1]

A major stumbling block is the cumbersome mechanism that must be invoked to explain a B-to-Z interconversion. As mentioned previously, a simple twisting of the helix from right to left is not sufficient, because the backbone chains run in opposite directions in the two forms. Fig. 23.4.3.6 demonstrates the steps that must still be undertaken after both B and Z helices have been unwound so as to remove all of their helical character. Note the opposite sense of the backbone strands in B [part (*a*)] and Z [part (*e*)]. In order to accomplish the interconversion, base pairs of B-DNA must be pulled apart, as in part (*b*), and each base pair swung around to the opposite side of the backbone 'ladder' [part (*c*)]. This would automatically lead to *syn* conformations at both ends of the base pair, as drawn in Fig. 23.4.3.4(*b*). Returning pyrimidines to an *anti* conformation would create the zigzag backbone chain (Fig. 23.4.3.4*c*). Base pairs can then be restacked, as in parts (*d*) and (*e*) in Fig. 23.4.3.6 (which differ only by rotation of the entire helix about the vertical), to yield the backbone geometry of a Z helix. This is the simplest interconversion and one which was recognized and proposed in the very first Z-DNA structure paper (Z1). Other alternatives have been suggested, involving breaking individual base pairs, swinging the bases independently around their backbone chains, and re-forming the pairs. But one kind of special mechanism or another must be invoked if a B-to-Z interconversion is to be achieved.

23.4.3.7. 'Watson–Crick' Z-DNA

Ansevin & Wang (1990) have proposed an alternative left-handed double helix, with many of the properties of Z-DNA, but possessing the same backbone chain orientations as A- and B-DNA. With such a helix, a B-to-Z conversion would require only a twisting of the duplex about its axis – no separation of bases or unpairing, and no pulling apart of the stack. Ansevin & Wang did not challenge the X-ray crystal structure analyses of short Z-DNA oligomers. Instead, they suggested that Z-DNA was globally the most stable form, adopted in short oligomers where chain unravelling and rearrangement is easy, but that their 'Watson–Crick' Z-DNA or Z(WC)-DNA was the structure that was actually produced by *in vitro* or *in vivo* manipulations of long DNA duplexes. They noted that most solution measurements focus on only two characteristics of the DNA: left-hand-

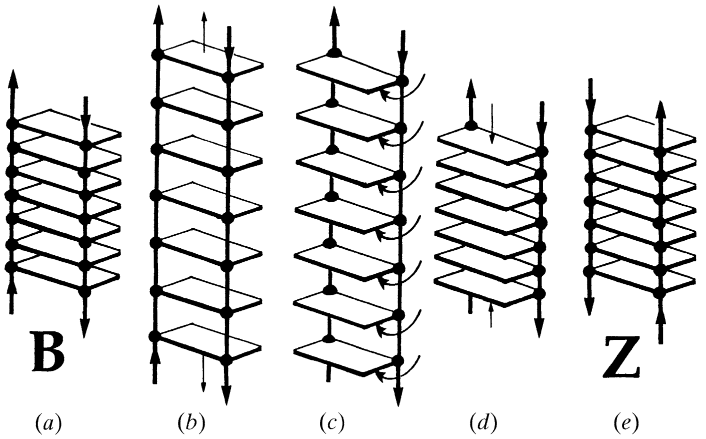


Figure 23.4.3.6
Interconversion of a B to a Z helix. Because the strands have opposite directions in B (*a*) and Z (*e*), interconversion must involve opening up the helix (*b*), flipping each base pair to the other side (*c*), and re-stacking base pairs (*d*). (*d*) and (*e*) are identical upon rotation about a vertical axis.

---

[1] Rich and co-workers (Schwartz *et al.*, 1999) have recently solved the crystal structure of the Zα domain of the human editing enzyme ADAR1 in a complex with a six-base-pair Z-DNA helix of sequence CGCGCG. This left-handed hexamer may suffer from the same length *versus* conformation uncertainty mentioned later in this chapter in connection with oligonucleotide crystals, especially since protein–DNA contacts in the Zα complex occur only with the zigzag phosphate backbone, which is not that dissimilar in Z-DNA (Fig. 23.4.3.3) and Z(WC)-DNA (Fig. 23.4.3.7). Nevertheless, it is encouraging to see a short segment of Z actually making contact with its protein in a presumably biologically relevant context.

edness and a dinucleotide repeat, both shared by Z-DNA and Z(WC)-DNA.

The Z(WC) helix is shown in Fig. 23.4.3.7, and a different stereo view appears as Fig. 7 of Dickerson (1992). Like Z-DNA, it is left-handed, with a deep minor groove and shallow major groove. Cytosines with *anti* glycosyl bonds and guanines with *syn* bonds alternate along each backbone strand. However, sugar puckering is reversed: cytosines are C3′-*endo*, while guanines are C2′-*endo*. In Z-DNA, the backbone chain runs parallel to the helix axis past G, and at right angles to the axis past C. In Z(WC)-DNA, this is reversed: parallel to the helix past C, and at right angles past G. Because of efficient stacking of base pairs, the logical two-base-pair structural unit in Z-DNA is $^{5'}$C–G$^{3'}$; in Z(WC)-DNA it is $^{5'}$G–C$^{3'}$. One such unit is clearly visible in the centre of Fig. 23.4.3.7. This behaviour is reflected in local twist angles:

Helix	C–G	G–C	Sum
Z-DNA	−8°	−52°	−60°
Z(WC)-DNA	−70°	+10°	−60°

The Ansevin–Wang helix has been sedulously ignored since its publication in 1990, especially by crystallographers. The Science Citation Index lists an average of *one* citation of their paper per year since publication, most commonly by spectroscopists. Ho & Mooers (1996) are almost alone among crystallographers in coupling the B-to-Z interconversion dilemma to the possible existence of a different kind of left-handed structure in long polynucleotides. Of course the Z(WC)-DNA structure, as presented here, is only a model; it could be far from the true structure in many respects. But its interest lies in the fact that a left-handed alternating helix with 'standard' backbone directions *can* be built with reasonable bond geometries and with properties that fit the various physical measurements as well as Z-DNA. It calls into question not the correctness of the Z-DNA structure obtained from short oligomers with free helix ends, but the relevance of that structure to the production of left-handed regions in longer duplexes with constrained ends.

23.4.4. Sequence–structure relationships in B-DNA

Two channels of information exist in B-DNA by which base sequence is expressed to the outside world. One of these is the Watson–Crick base pairing of A with T and G with C that is used in the storage of genetic information and in replication and transcription. The other channel, used in control and regulation of the expression of this genetic information, involves the hydrogen-bonding patterns of base-pair edges along the floors of the grooves and any systematic deformations of local helix structure that result explicitly from the base sequence.

The simplest and most direct expression of this second channel is the passive reading of hydrogen-bonding patterns along the floor of the major and minor grooves. This readout mechanism was first proposed by Seeman *et al.* (1976), and involves acceptors and donors as marked by *A* and *D* in Fig. 23.4.2.7. The wide major groove of B-DNA is read by several classes of control proteins that function by positioning an α-helix within the groove so that its amino-acid side chains can sense the pattern of hydrogen bonding. This category includes prokaryotic and eukaryotic helix-turn-helix or HTH proteins, zinc-finger and other zinc-binding proteins, basic leucine zippers and their basic helix-loop-helix cousins, and others (See Table I of Dickerson & Chiu, 1997). The narrower minor groove is a frequent target for long, planar drug molecules, such as netropsin and distamycin, as listed in Part II of Table A23.4.1.2.

In principle, this readout mechanism would work perfectly well with a regular, ideal, fibre-like B-DNA helix. But other control proteins that recognize the minor groove, such as TATA-binding protein (TBP) and integration host factor (IHF), depend not merely on passive hydrogen bonding to an ideally regular duplex, but on the *sequence-dependent deformability* of one region of the helix *versus* another. The remainder of this chapter will be concerned with this effect and its role in DNA recognition.

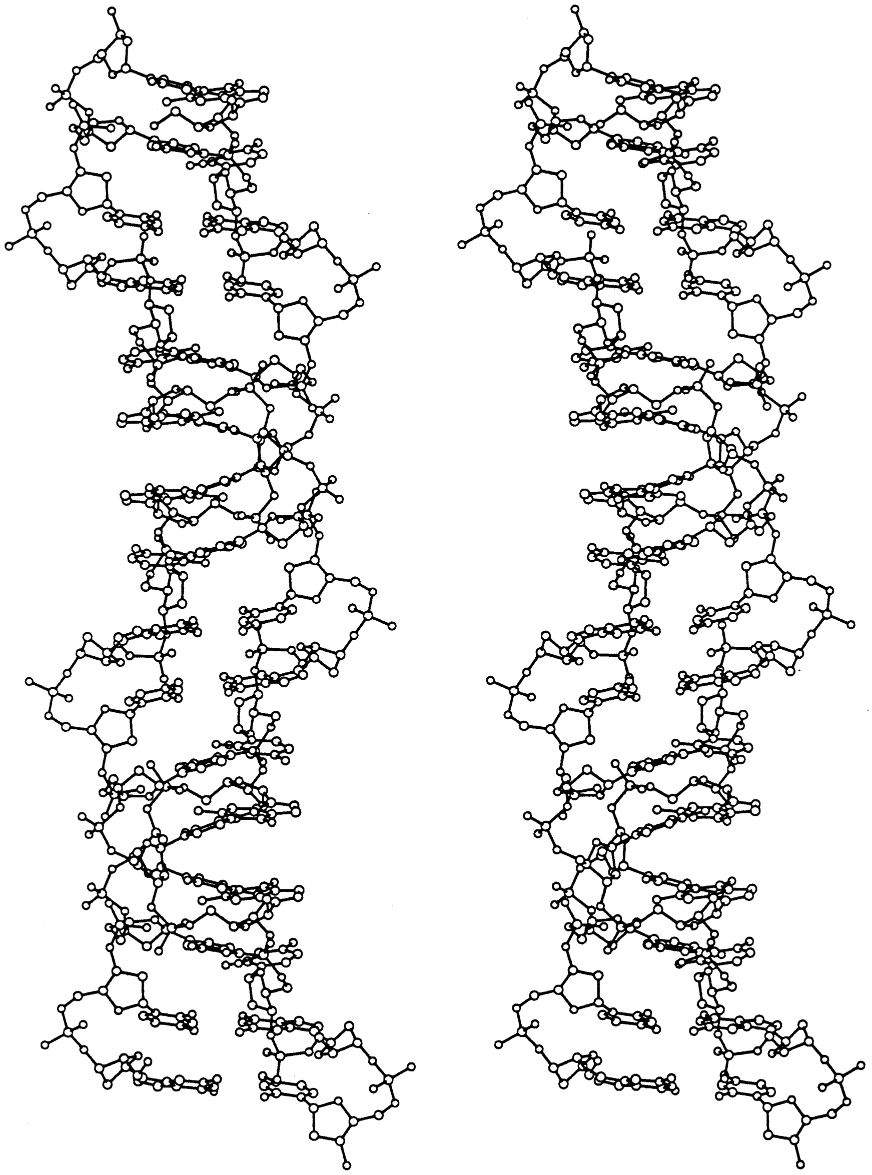


Figure 23.4.3.7
Z(WC)-DNA, or 'Watson–Crick Z-DNA', a proposed left-handed, zigzag, alternating purine/pyrimidine helix with many of the properties of Z-DNA, but with the backbone chain sense found in A- and B-DNA (Ansevin & Wang, 1990). Coordinates courtesy of Allen T. Ansevin.

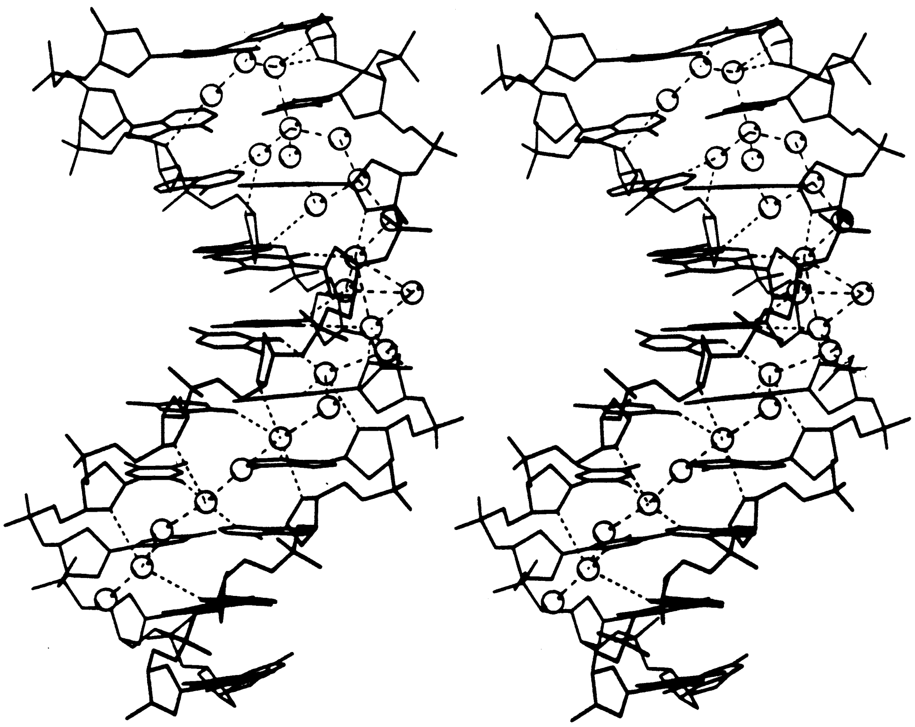


Figure 23.4.4.1
Structure of C-A-A-A-G-A-A-A-A-G (B107). The lower half of the helix, with -A-A-A-A-G, exhibits the narrow minor groove commonly associated with the AT region of the helix and a single zigzag spine of hydration, as was first seen in C-G-C-G-A-A-T-T-C-G-C-G (B1–B6). The upper half, with C-A-A-A-G-, has the wider minor groove of general-sequence B-DNA and two separate rows of hydrating water molecules along the two walls of the wider groove.

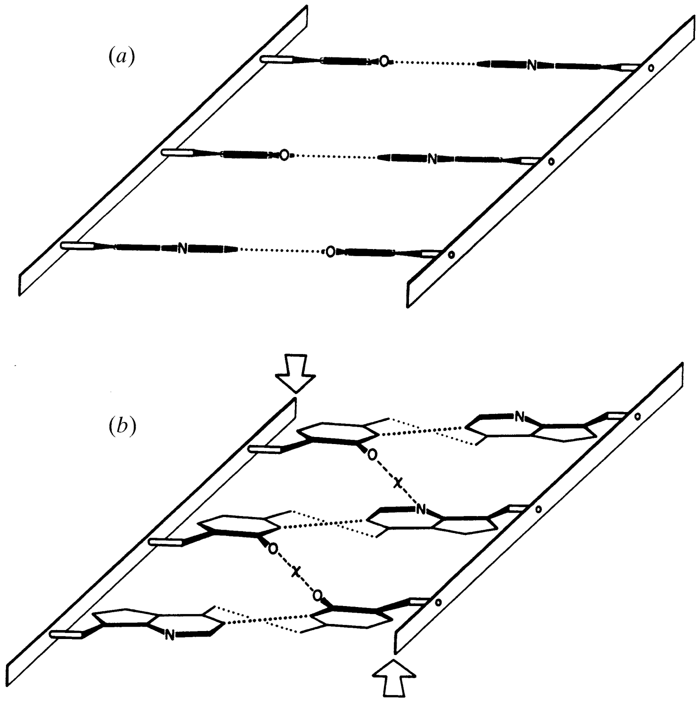


Figure 23.4.4.2
Relationship between minor groove width and propeller twist. (*a*) View into the minor groove of B-DNA, with base pairs seen on edge and with the sugar–phosphate backbones shown schematically as inclined ladder uprights. (*b*) Consequences of propeller twisting the base pairs. Glycosyl bonds connected to sugar C1′ atoms are all displaced upward in the right strand and downward in the left strand. This shifts the backbone chains as indicated by the arrows. Hence, the gap between the chains is decreased, and the minor groove is narrowed.

23.4.4.1. Sequence-dependent deformability

23.4.4.1.1. Minor groove width

The simplest and first-noticed sequence-dependent deformability of the B-DNA duplex was variation in minor groove width. The first B-DNA oligomer to be solved, C-G-C-G-A-A-T-T-C-G-C-G (B1–B6), had a narrow minor groove in the central A-A-T-T region, with only *ca* 3.5 Å of free space between opposing phosphates and sugar rings. (It has become conventional to define the free space between phosphates as the measured minimal P–P separation across the groove, less 5.8 Å to represent two phosphate-group radii. Similarly, the measured distance between sugar oxygens is decreased by 2.8 Å, representing two oxygen van der Waals radii.) The C-G-C-G ends of the helix had the 6–7 Å opening expected for ideal B-DNA, but the situation was clouded, because the outermost two base pairs at each end of the helix interlocked minor grooves with neighbours in the crystal. Hence, the wider ends could possibly be only an artifact of crystal packing.

After 1991, the situation was clarified by the structures of several decamers [Table A23.4.1.2, Part I(*c*)], which stack on top of one another without the interlocking of grooves. The normal minor groove opening is *ca* 7 Å. Regions of four or more AT base pairs can exhibit a significantly narrowed minor groove, although such narrowing is not mandatory. This behaviour is seen with the B-DNA decamer, C-A-A-A-G-A-A-A-A-G, in Fig. 23.4.4.1. The narrowing arises mainly from the larger allowable propeller twist in AT base pairs, which displaces C1′ atoms at opposite ends of the pair in different directions, and moves the backbone chains in such a way as to partially close the groove (Fig. 23.4.4.2).

This is an excellent example of the concept of *sequence-dependent helix deformability*, rather than simple deformation. The two hydrogen bonds of an AT base pair allow a larger propeller twist but do not require it. Hence, AT regions of helix permit a narrowing of the minor groove but do not demand it. Indeed, this lesson was brought home in the most dramatic way when Pelton & Wemmer (1989, 1990) showed *via* NMR that a 2:1 complex of distamycin with C-G-C-A-A-A-T-T-G-G-C or C-G-C-A-A-A-T-T-T-G-C-G could exist, in which two drug molecules sat side-by-side within an enlarged central minor groove. Fig. 23.4.4.3 shows a narrow minor groove with a single netropsin molecule, and Fig. 23.4.4.4 shows a wide minor groove enclosing two di-imidazole lexitropsins side-by-side. In summary, an AT-rich region of minor groove is capable of narrowing but is not inevitably narrow, in contrast to GC-rich regions where the third hydrogen bond tends to keep the base pairs flat and the minor groove wide. The AT minor groove is potentially *deformable* without being inevitably *deformed*.

23.4.4.1.2. Helix bending

Sequence-dependent bendability has been reviewed recently by Dickerson (1988*a*,*b*,*c*) and Dickerson & Chiu (1997). The relative bendability of different regions of B-DNA sequence is an important aspect of recognition, one that is used by countless control proteins that must bind to a particular region of double helix. Catabolite activator protein or CAP (Schultz *et al.*, 1991; Parkinson *et al.*, 1996), *lacI* (Lewis *et al.*, 1996) and *purR* (Schumacher *et al.*, 1994) repressors, $\gamma\delta$-resolvase (Yang & Steitz, 1995), *Eco*RV restriction enzyme (Winkler *et al.*, 1993; Kostrewa & Winkler, 1995), integration host factor or IHF (Rice *et al.*, 1996), and TBP or TATA-binding protein (Kim, Gerger *et al.*, 1993; Kim, Nikolov & Burley, 1993; Nikolov *et al.*, 1996; Juo *et al.*, 1996) are all sequence-specific DNA-binding proteins that bend or deform the nucleic acid duplex severely during the recognition process. IHF in Fig. 23.4.4.5 may be taken as representative of this class of DNA-binding proteins. The bend is produced by two localized rolls of *ca* 60° in a direction compressing the major groove and are additive, because they are spaced nine base pairs, or roughly one turn of helix, apart. In IHF, the two helix segments flanking the bend should be straight and unbent, and this is accomplished in one segment *via* a six-adenine A-tract: -C-A-A-A-A-A-A-G-.

The bending locus in IHF is C-A-A-T/A-T-T-G . It is C-G in *lacI* and *purR* repressors (Fig. 23.4.4.6), C-A = T-G in CAP (Fig. 10 of Dickerson, 1998*b*), and T-A in *Eco*RV, $\gamma\delta$-resolvase and TBP (Fig. 23.4.4.7). Pyrimidine-purine or Y-R steps appear to be especially suitable loci for roll bending. The dashed lines in Figs. 23.4.4.6 and 23.4.4.7 plot tilt, and demonstrate its insignificance in bending, compared with roll. (This is intuitively obvious. Imagine yourself standing near a tall stack of wooden planks in a lumberyard during an earthquake. Where would you prefer to stand: alongside the stack, or at one end?)

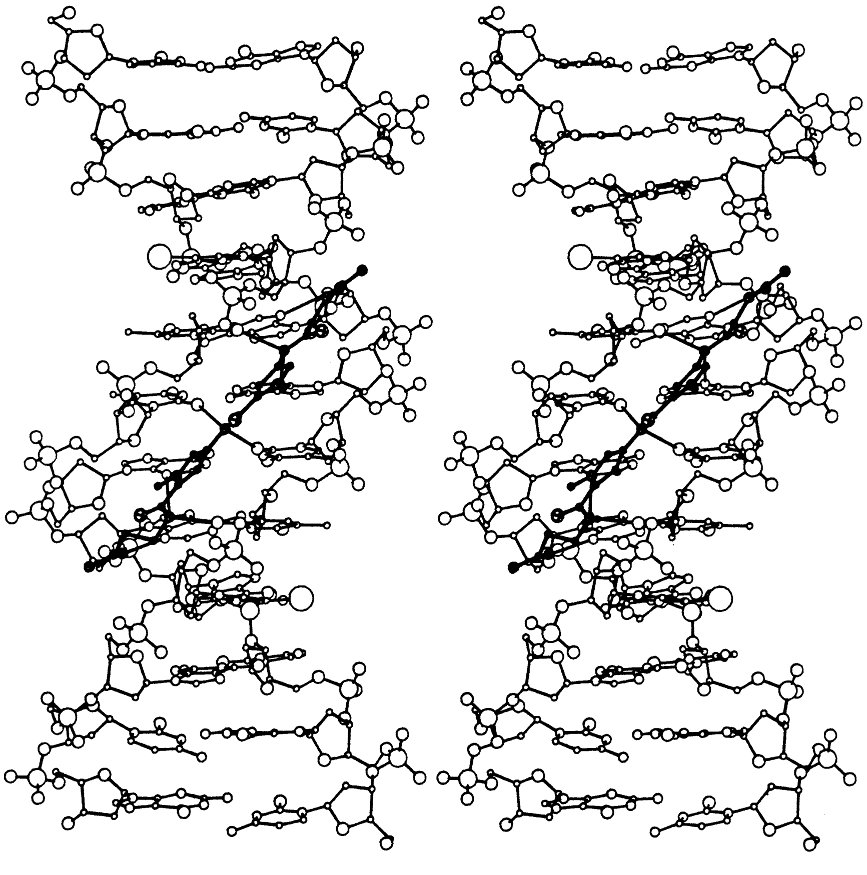


Figure 23.4.4.3
Structure of the 1:1 complex of netropsin with C-G-C-G-A-A-T-T-C-G-C-G (B11, B12, B87). The drug binds to the central -A-A-T-T- region of the minor groove, which is barely wide enough to enclose the nearly planar polyamide molecule. The netropsin structure can be represented by

$$^{+}(NH_2)_2C{-}NH{-}CH_2{-}CONH{-}Py{-}CONH{-}Py{-}CONH{-}CH_2{-}CH_2{-}C(NH_2)_2^{+}$$

where Py is a five-membered methylpyrrole ring. An even more compact representation, useful when comparing other polyamide netropsin analogues or lexitropsins, is $^{+}$=Py=Py=$^{+}$, where the common cationic tails are indicated only by a plus sign, and = represents a —CONH— amide.

In summary, bending of the B-DNA helix nearly always involves roll, not tilt. The easier direction of bending is that which compresses the broad major groove, although examples of roll compression of the minor groove are known. Y-R steps are especially prone to roll bending. Again, the phenomenon is one of sequence-induced bendability, not mandatory bending. No one imagines that the IHF binding sequence of Fig. 23.4.4.5 is permanently kinked at its two C-A-A-T/A-T-T-G steps, wandering deformed through the nucleus, looking for an IHF molecule to bind to. Instead, this sequence has a potential bendability that other sequences, such as A-A-A-A-A-A, lack.

Table 23.4.4.1 summarizes the observed behaviour of Y-R, R-R and R-Y steps from a great many X-ray crystal structure analyses, with and without bound DNA. In the present context, these rules are termed the ‘Major Canon’, since they are well established and generally well understood. Some understanding of the proneness of Y-R steps to bend can be obtained by looking at stereo pairs of two successive base pairs viewed down the helix axis. Fig. 23.4.4.8 gives a few representative examples; many more can be found in Figs. 4–6 of Dickerson (1988*b*) and in the original literature. In brief, Y-R steps, especially C-A and T-A, tend to orient so that polar exocyclic N and O atoms stack against polarizable rings of the other base pair. This is the same type of polar-on-polarizable stacking stabilization mentioned earlier in connection with O4′ and guanine in Z-DNA (Bugg *et al.*, 1971; Thomas *et al.*, 1982; Hunter &

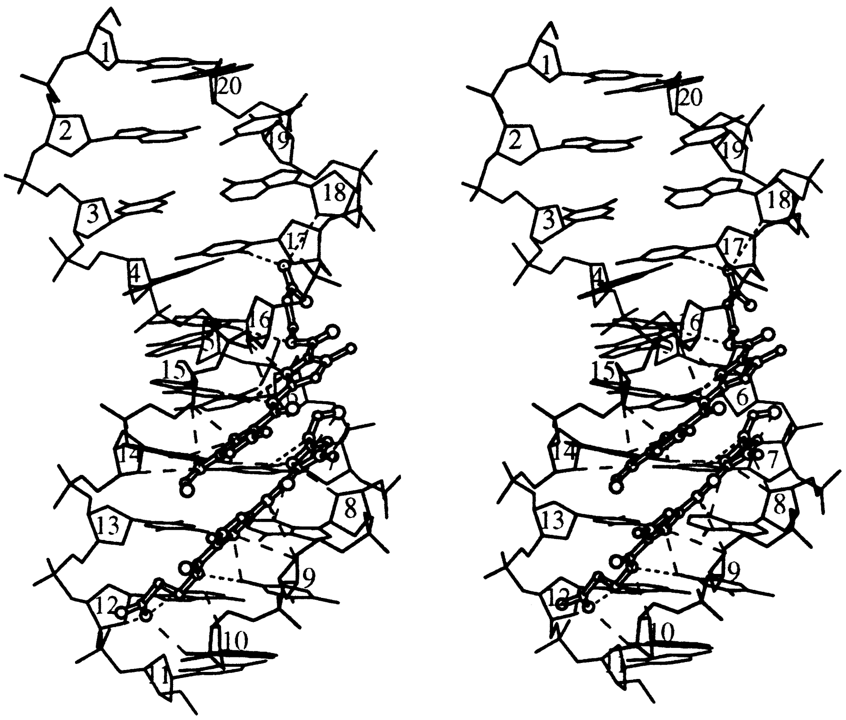


Figure 23.4.4.4

Structure of the 2:1 complex of a di-imidazole lexitropsin with C-A-T-G-G-C-C-A-T-G (B108). The drug now is represented by

$$\text{H–CONH–Im–CONH–Im–CONH–CH}_2\text{–CH}_2\text{–C(NH}_2)_2^+$$

where Im is a five-membered imidazole ring, or again more compactly by ^{0}=Im=Im=$^+$. The uncharged leading amide group, characteristic of distamycins, is identified by ^{0}. Distamycin itself would be represented in this shorthand notation by ^{0}=Py=Py=Py=$^+$. Reprinted from B108, copyright (1977), with permission from Excerpta Medica Inc.

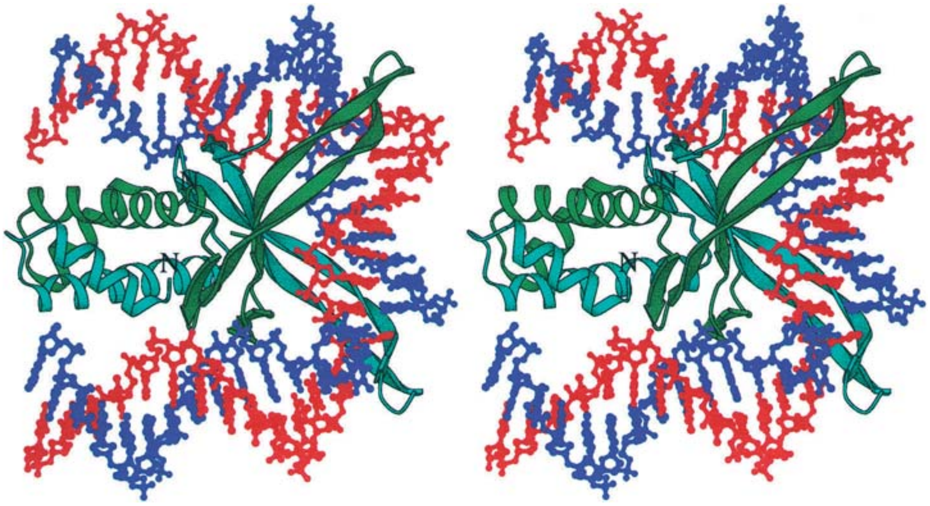


Figure 23.4.4.5

DNA duplex (red and blue strands) looped around IHF or integration host factor. The two subunits of the IHF duplex are green and turquoise. Two antiparallel loops of protein chain, one from each subunit, insert into the minor groove of B-DNA at the sequence C-A-A-T/A-T-T-G and produce abrupt bends *via* local roll angles of 60°. The two localized bends are additive because they occur one helical turn apart. All other steps have roll angles of 5° or less. The two flanking helix segments pack against the IHF dimer and must be kept straight and unbent. This is accomplished in one of the two segments by an A-tract of sequence C-A-A-A-A-A-A-G. From Dickerson & Chiu (1997). Coordinates courtesy of P. Rice.

Sanders, 1990; B32). Base pairs in T-A steps tend not to slide over one another along their long axes, keeping pyrimidine O2 stacked over the purine five-membered ring (Fig. 23.4.4.8*b*). C-A steps can adopt this same stacking, or the base pairs can slide until the pyrimidine O2 sits over the purine six-membered ring instead (Fig. 23.4.4.8*a*).

Purine-purine or R-R steps behave quite differently (Fig. 23.4.4.8*c*). They stack ring-on-ring, usually with greater overlap on the purine end than the pyrimidine. The net effect is that the pivot appears to pass through or near the purines, while pyrimidines at the other end of the pairs stack O2-on-ring as with Y-R steps. R-Y steps tend to stack ring-on-ring, with little contribution from exocyclic atoms.

El Hassan & Calladine (1997) have recently examined roll, slide and twist behaviour at 400 different steps observed in crystal structures of 24 A- and 36 B-DNA oligomers. The author has carried out a similar analysis of 1137 steps from 86 sequence-specific protein–DNA complexes (Dickerson, 1998*a*,*c*; Dickerson & Chiu, 1997). A striking feature is that trends in local parameters are just the same in DNA crystals and in protein–DNA

Table 23.4.4.1

Sequence-dependent differential deformability in B-DNA. I. The Major Canon

See Dickerson (1998*a*,*b*,*c*) and Dickerson & Chiu (1997).

(1) *Structural basis for helix bending in B-DNA*

Bending is nearly always the result of roll between successive base pairs, seldom tilt.

Positive roll, compressing the wide major groove, is more common than negative roll, in which the narrower minor groove is compressed.

Observed bends in B-DNA are of three main types: (*a*) *localized kinks* (large positive roll at one or two discrete base steps), (*b*) *three-dimensional writhe* (positive roll at a series of successive steps), or (*c*) *smooth curvature* (alternation of positive and negative roll every half turn, with side-to-side zigzagging at intermediate positions). (*a*) and (*b*) are easier to accomplish than (*c*), and hence are more common.

Local writhe in a DNA helix produces macroscopic curvature only when the extent of writhe does not match the natural rotational periodicity of the helix. Endless writhe results in a straight helix, and indeed A-DNA can be regarded as a continuously writhed variant of the B form. Conversely, the bending effect of writhe can be amplified if it is repeated with the periodicity of the helix itself – that is, repeated alternation of writhed and unwrithed segments every ten base pairs, as with A-tract B-DNA.

(2) *Pyrimidine-purine (Y-R) steps: C-A = T-G, T-A and C-G*

Little ring–ring stacking overlap.

Polar N or O stacked over polarizable aromatic rings.

Y-R steps are natural fracture points for the helix. They can show (but are not required in every case to show) large twist and slide deformations, and bending mainly *via* positive roll, compressing the major groove.

(3) *Purine-purine (R-R) steps: A-A = T-T, A-G = C-T, G-A = T-C and G-G = C-C*

Extensive ring–ring overlap.

Base pairs tend to pivot about stacked purines as a hinge, with greater ring–ring separation at pyrimidine ends.

Tight stacking, with only minor roll, slide and twist deformations.

(4) *Purine-pyrimidine (R-Y) steps: A-C = G-T, A-T and C-G*

Behaviour in general like R-R steps, with extensive ring–ring overlap and tight stacking, with again only minor roll, slide and twist deformations.

(5) *A-A and A-T steps, as contrasted with T-A*

Especially resistant to roll bending, probably because of sawhorse interlocking of highly propellered base pairs, supplemented by inter-base-pair hydrogen bonds within grooves. In contrast, T-A is particularly weak and subject to roll bending.

A-tracts, defined as four or more consecutive AT base pairs without the disruptive T-A step, are especially straight and resistant to bending. Natural selection has apparently chosen short A-tracts for regions of protein–DNA contacts where bending is not wanted.

complexes. The frequently invoked nightmare of 'crystal packing deformations' appears to be of only minor significance. In both studies (El Hassan & Calladine, 1997; Dickerson, 1998*b*), roll *versus* slide, slide *versus* twist and twist *versus* roll plots are presented for all ten possible base-pair steps. Fig. 23.4.4.9 illustrates roll *versus* slide plots for two Y-R, two R-R and two R-Y steps.

Table 23.4.4.2 summarizes observations from these roll/slide/twist plots. These are labelled the 'Minor Canon' since they are recent, approximate and not well understood. However, they provide goals for future investigations of helix behaviour.

23.4.4.2. *A-tract bending*

It has long been known that introduction of short A-tracts into general-sequence B-DNA in phase with the natural 10–10.5 base-pair repeat produced overall curvature that could be detected *via* eletrophoretic gel retardation, ring-cyclization kinetics and other physical measurements in solution (Marini *et al.*, 1982; Wu & Crothers, 1984; Koo *et al.*, 1986; Crothers & Drak, 1992). However, the microscopic source of the observed macroscopic curvature remained unclear. Solution measurements alone cannot discriminate between three alternative curvature models: (1) local bending within the A-tracts themselves; (2) bending at junctions between A-tract B-DNA and general-sequence B-DNA; or (3) inherently straight and unbent A-tracts, with curvature resulting from removal of the normal writhe expected in general-sequence B-DNA (Koo *et al.*, 1990; Crothers *et al.*, 1990). The three curvature models are compared schematically in Fig. 10 of reference B77.

X-ray crystallographic results for DNA oligomers come down unequivocally in favour of model (3) above. Short A-tracts of four to six base pairs are straight and unbent in C-G-C-G-A-A-T-T-C-G-C-G (B1–B6), C-G-C-A-A-A-A-A-A-G-C-G (B20), C-G-C-A-A-A-A-A-T-G-C-G (B31), C-G-C-A-A-A-T-T-T-G-C-G (B17, B52), C-G-C-G-A-A-A-A-A-A-G-C (B64) and C-A-A-A-G -A-A-A-A-G (B105) (A-tracts are double-underlined). It has been claimed (Sprous *et al.*, 1995) and disputed (Dickerson *et al.*, 1994, 1996) that the observed straightness of crystalline A-tracts was only an artifact of crystal packing, or of the high levels of methyl-2,4-pentanediol (MPD) used in the crystallization. This concern now is put to rest by the observation that B-DNA packed against a protein molecule in its biological working environment behaves exactly the same as B-DNA packed against other DNA molecules in the crystal, as borne out by the roll/slide/twist studies of El Hassan & Calladine (1997) for DNA and of Dickerson (1998*a*,*b*,*c*) and Dickerson & Chiu (1997) for protein–DNA complexes. Added support has come from recent molecular-dynamics simulations by Beveridge and co-workers (Sprous *et al.*, 1999), who have demonstrated that the duplex of sequence GGGGGGAAAATTTTCGAAAATTTTCCCCCC is severely curved because of a roll kink at the double-underlined central CG step, whereas the duplex GGGGGGTTTTAAAACGTTTTAAAACCCCCC is much less curved because the roll kink at CG is counterbalanced by roll kinks in the opposite direction at the two flanking TA steps. In both cases, A-tracts are straight and completely unbent. (Note that both roll kinks can involve compression of the major groove, as expected, because the kink sites are a half turn of helix apart.)

This similarity of behaviour of DNA in crystals and in protein–DNA complexes should come as no surprise, since the local molecular environments – close intermolecular contacts, partial dehydration, low water activity, low local dielectric constant, high ionic strength, presence of divalent cations – are similar in these two cases and quite different from that of free DNA in dilute

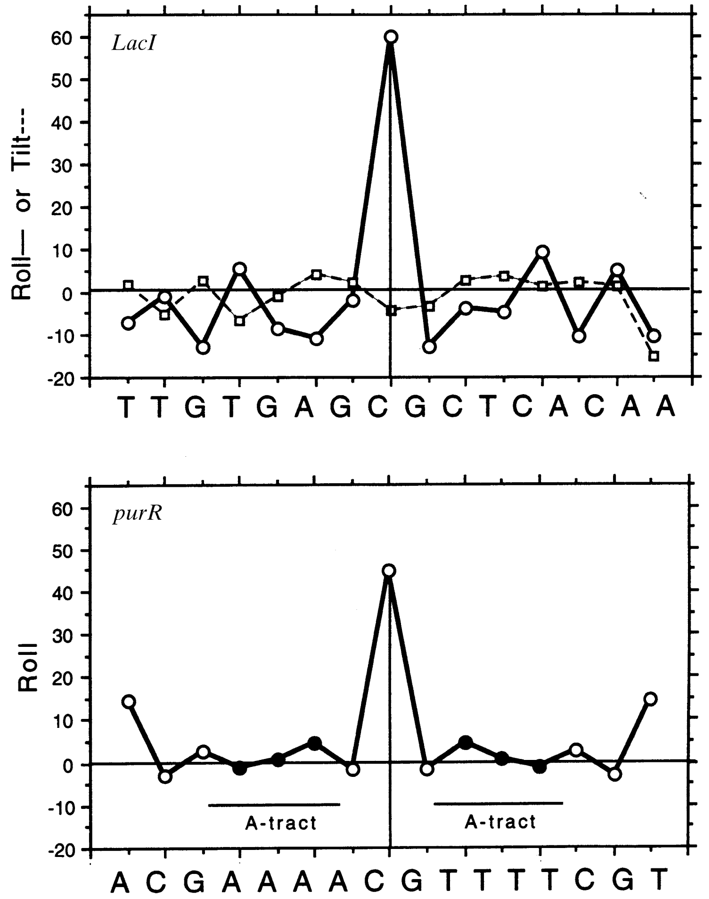


Figure 23.4.4.6
Roll-angle plots for sequence-specific DNA–protein complexes with *lacI* (top) and *purR* (bottom). In each case, bending occurs *via* localized roll at a C-G step. Other steps of the sequence have random rolls of *ca* 10° or less. Note that, as with IHF, A-tracts are especially straight and unbent. Dashed lines in the *lacI* plot demonstrate the unimportance of tilt in production of helix bending.

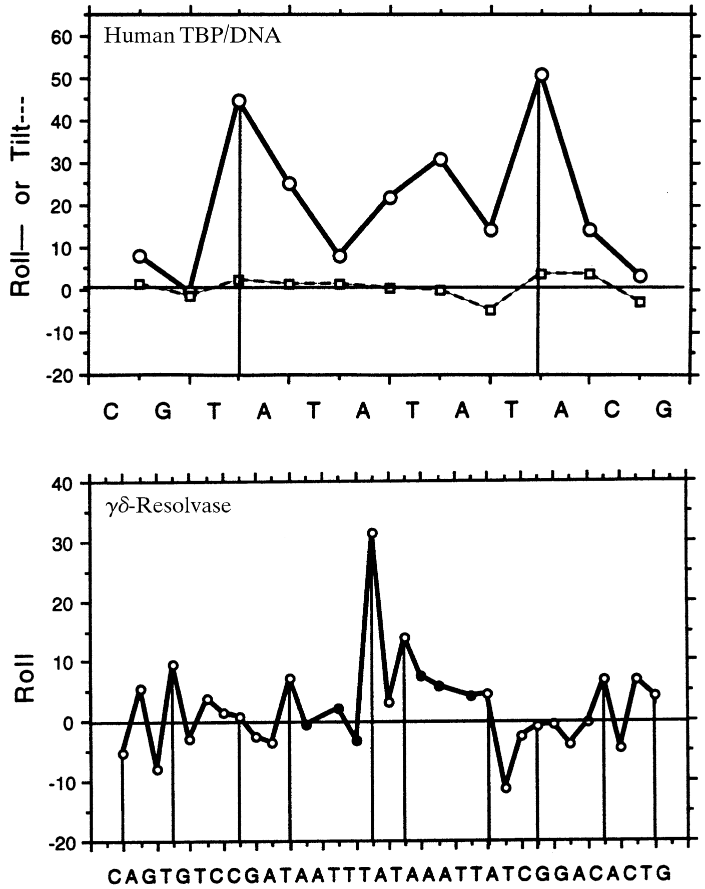


Figure 23.4.4.7
Bending *via* roll at T-A steps in TBP or the TATA-binding protein (top) and in γδ-resolvase (bottom). Note that not every T-A step in TBP or γδ-resolvase is necessarily bent. Note also in γδ-resolvase that C-A = T-G steps, which in proteins such as CAP are used to generate sharp roll bends, here, frequently, are local roll maxima, even though they contribute little to the overall bending. They have a bending *potential* that is not used in this particular setting.

aqueous solution. Far from being unwanted 'crystal deformations', the local changes in structure resulting from intermolecular contacts in DNA crystals provide positive information about sequence-dependent deformability that is relevant to the protein recognition process. With regard specifically to A-tract behaviour, Occam's Razor would argue in favour of model (3) above for the behaviour of A-tracts in solution. The situation in dilute aqueous solution becomes of secondary importance if what is wanted is an understanding of A-tract B-DNA behaviour in protein–DNA complexes. Here, the answer is unambiguous: A-tracts in their biological setting are inherently rigid structural elements, chosen by natural selection when bending should be avoided.

23.4.5. Summary

Three families of nucleic acid double helix have been found – A, B and Z – with widely different structures and usages. The A and B helices are right-handed and have no limitations on base sequence. Z is left-handed and effectively limited to alternating purines and pyrimidines, with G and C overwhelmingly favoured. B is the biologically significant helix for DNA and is used in genetic coding. A is the helix of preference for RNA because it can accommodate the C2′-OH group of ribose, which produces steric clash in the B helix. The Z helix has, as yet, no well established biological function. A left-handed DNA configuration can be induced in longer DNA segments by negative supercoiling in solution, but it is not clear that this left-handed configuration is identical to the Z-DNA seen in short crystalline oligomers, because of the reversed orientation of backbone strands in Z-DNA.

B-DNA is an inherently malleable or deformable duplex. Its sugar ring conformations are much more variable than those of A-DNA. The base sequence of B-DNA is expressed directly *via* hydrogen bonds between bases of a pair, and indirectly *via* hydrogen-bond donors and acceptors along the floor of the major and minor groove. Sequence is also expressed as a *differential deformability* of different regions of the duplex. The two most obvious parameters affected by base sequence are minor groove width and helix bendability. Certain sequences of B-DNA are not statically bent, but are more bendable under stress than are other sequences. Bending occurs *via* roll, usually in the direction that compresses the broad major groove. Pyrimidine-purine or Y-R steps are most conducive to roll bending, and purine-purine steps are least bendable, particularly A-tracts of four or more AT base pairs without the weak T-A step. Natural selection has engineered Y-R steps into a DNA sequence where a sharp roll bend is wanted, and short A-tracts into a sequence where bending is not desired.

Table 23.4.4.2

Sequence-dependent differential deformability in B-DNA. II. The Minor Canon

These generalizations are illustrated by Fig. 23.4.4.9, and are justified at greater length by El Hassan & Calladine (1997) and Dickerson (1998*a*,*b*,*c*).

(6) *Heterogeneous steps ending in A: C-A, T-A and G-A*

Steps ending in adenine, aside from A-A, tend to display (*a*) negative correlation between slide and roll, and between twist and roll, and (*b*) positive correlation between slide and twist.

(7) *Purine-pyrimidine steps*

R-Y steps display, on average, a systematic preference for negative slide and for twist below 36°.

(8) *Relative step frequencies in sequence-specific protein–DNA complexes*

Step A-A is the most common of all, and in 55% of the cases it occurs within A-tracts.

Steps containing only GC base pairs are least common, and seemingly are less compatible with formation of sequence-specific protein complexes.

(9) *Local environment and DNA behaviour*

Sequence-dependent local helix deformations are quite similar in DNA crystals and in protein–DNA complexes. DNA molecules packed against proteins in their normal biological environment appear to have more in common with DNA packed against other DNA helices in the crystal than with free DNA in solution.

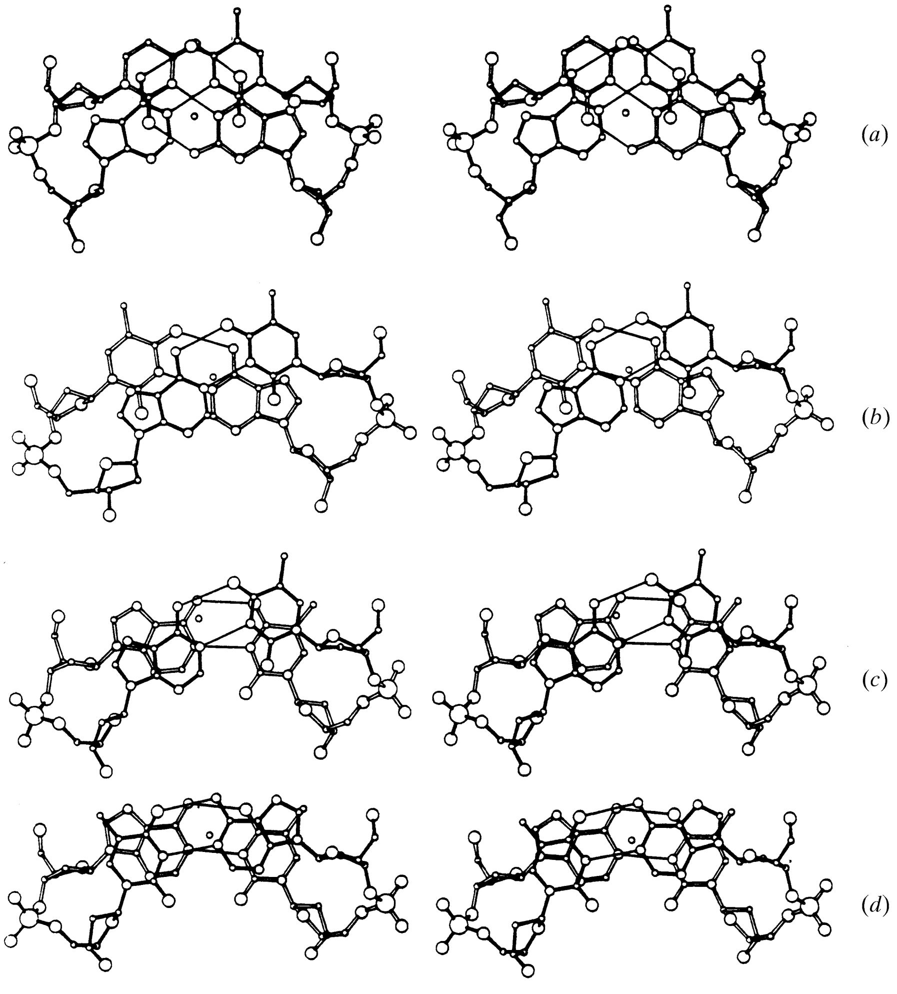


Figure 23.4.4.8

Representative base-pair steps from B-DNA single-crystal X-ray analyses. (*a*) Pyrimidine-purine C-A step from C-C-A-A-G-A-T-T-G-G (B22, B46) (roll/slide/twist = −7.4°/2.6 Å/49.9°). Note the lack of ring-on-ring stacking, replaced by the stacking of pyrimidine O2 and purine N6 or O6, on aromatic rings of the adjacent base pair. This stacking opens up the twist angle to an unusual 50°. Note also the large +2.6 Å slide, which positions pyrimidine O2 over the six-membered rings of the neighbouring purines. (*b*) Pyrimidine-purine T-A step from C-G-A-T-A-T-A-T-C-G (B62) (roll/side/twist = 3.8°/−0.2 Å/39.5°). The stacking is similar to C-A, except that a near-zero slide positions pyrimidine O2 over the five-membered rings of purines. (*c*) Purine-purine A-A step from C-C-A-A-C-G-T-T-G-G (B46, B50) (roll/slide/twist = 8.8°/0.5 Å/28.7°). Ring-on-ring overlap now predominates, with consequently lowered twist angle and essentially zero slide. Note that purines are more extensively stacked than pyrimidines, which appear to be approaching the O2-on-ring stacking of Y-R steps. (*d*) Purine-pyrimidine A-T step from C-G-A-T-A-T-A-T-C-G (B62) (roll/slide/twist = 5.2°/0.0 Å/25.2°). Ring-on-ring stacking again lowers the twist angle and keeps slide around zero. Now there is no stacking of exocyclic N or O on neighbouring rings.

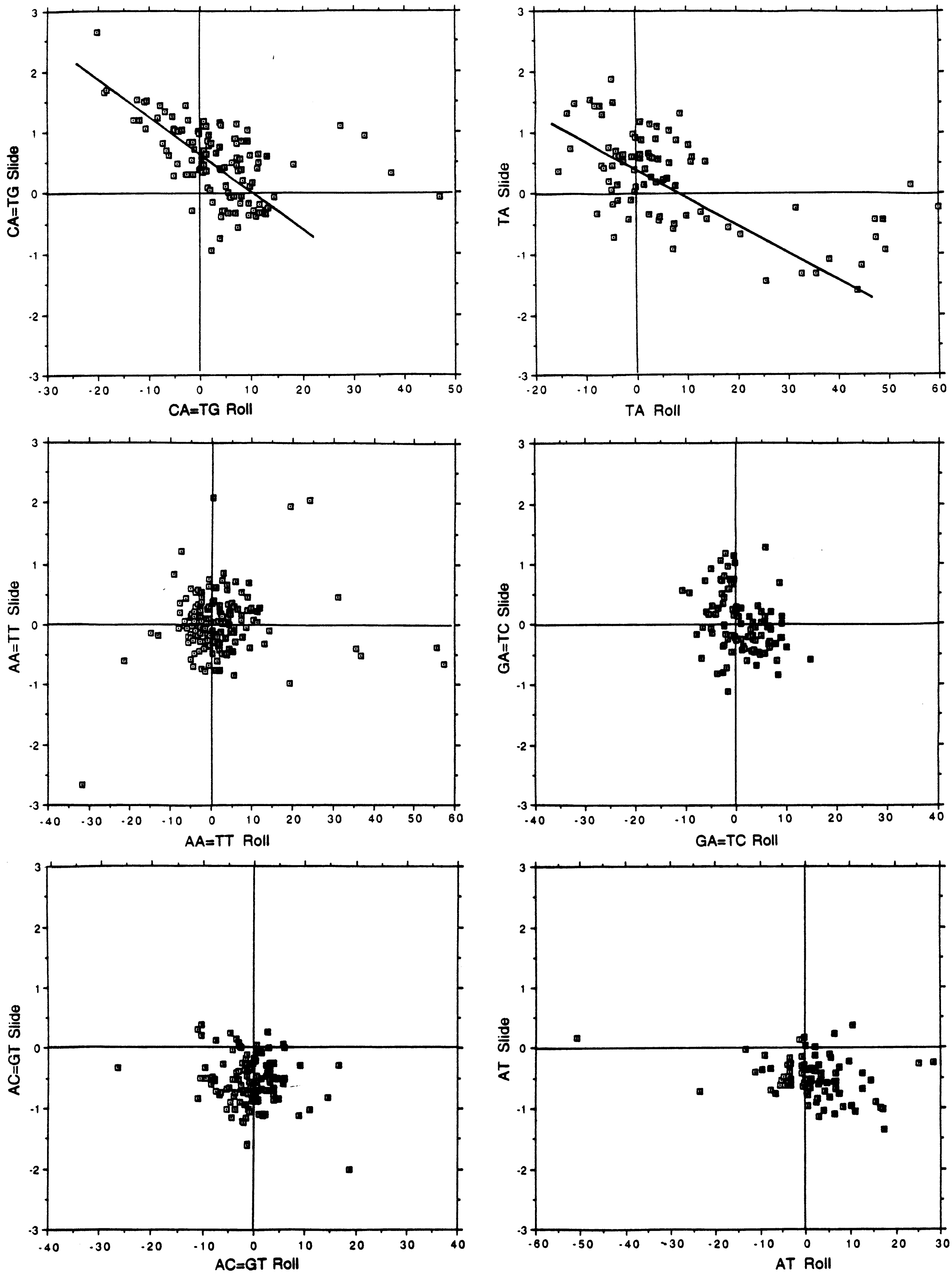


Figure 23.4.4.9

Slide *versus* roll plots for six of the ten possible base-pair steps. Data points are from 971 steps in crystal structure analyses of 63 sequence-specific protein–DNA complexes. A complete set of 30 plots for slide/roll, twist/roll and slide/twist at all ten steps is to be found in Dickerson (1998*b*), and equivalent plots for DNA alone are given by El Hassan & Calladine (1997). Y-R steps exhibit a broad range of roll, slide and twist values, with roughly linear correlations between pairs of variables. Points for A-A and other R-R steps cluster tightly around the origin, showing little tendency toward roll bending. Curiously enough, points for R-Y steps tend to favour negative values of slide and twist, and, hence, to concentrate in the lower left quadrant of a slide/tilt plot.

APPENDIX A23.4.1
X-ray analyses of A, B and Z helices

Table A23.4.1.1
X-ray analyses of A helices, DNA and RNA

This table and the two that follow are intended as a historical background and a focus on the geometry of the intact double helix. References are current as of late 1997; sequences marked 'to be published' in 1997 that still are unpublished two years later have been deleted. Also omitted are sequences with fewer than four base pairs in the asymmetric unit, complexes with intercalating drugs, helices with bulges or looped-out bases, unusual structures such as quadruplexes, hammerhead ribozymes and tRNA. For information on these and for more recent results, consult the Nucleic Acid Database (NDB) at http://ndbserver.rutgers.edu/. An NDB number in parentheses indicates that the authors have never made coordinates available to the public. These structures are of little scientific value, but have been included for historical reasons. Notes: Overhanging, unpaired bases are double underlined. Single underlining calls attention to mismatched bases or other interesting or relevant sequence aspects. Z = number of asymmetric units per cell. Ubp = number of base pairs per asymmetric unit. NDB No. = Nucleic Acid Database serial number. Abbreviations: 2am = 2-amino; 5br = 5-bromo; 6ame = 6′-α-methyl; 4mo = 4-methoxy; 5me = 5-methyl; 6aOH = 6′-α-hydroxyl; 6mo = 6-methoxy; 8oxo = 8-oxo; 6et = 6-ethyl; ara = arabinosyl; ps = phosphorothioate; (P) = leading phosphate; A, T, G, C = DNA; a, u, g, c = RNA; Py = pyrrole; Im = imidazole.

(*a*) Dodecamers

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
CCCCCGCGGGGG	$P3_221$	6	12	1991, Barcelona	ADL025	(A38)
CCGTACGTACGG	$P6_122$	12	6	1992, Ohio State	ADL045	(A41)
GCGTACGTACGC	$P6_122$	12	6	1992, Ohio State	ADL046	(A39)

(*b*) Decamers

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
GCGGGCCCGC	$P6_122$	12	5	1993, Ohio State	ADJ051	(A46)
GCACGCGTGC	$P6_122$	12	5	1996, Ohio State	ADJ075	(A60)
ACCGGCCGGT	$P6_122$	12	5	1989, MIT	ADJ022	(A26)
ACCGGCCGGT	$P6_122$	12	5	1995, MIT	ADJ065	(A55)
ACCCGCGGGT	$P6_122$	12	5	1995, MIT	ADJ066	(A55)
CCCGGCCGGG	$P2_12_12_1$	4	10	1993, Ohio State	ADJ049	(A47)
CCIGGCC^{5me}CGG	$P2_12_12_1$	4	10	1995, Ohio State	ADJB61	(A58)
GCGGGCCCGC	$P2_12_12_1$	4	10	1993, Ohio State	ADJ050	(A46)
ACCGGCCGGT	$P2_12_12_1$	4	10	1995, MIT	ADJ067	(A55)
CCGGGCCGCG	$P2_12_12_1$	4	10	1997, Ohio State	ADJ081,2	(A71)
C^{5me}CGGGCCCGG	$P2_12_12_1$	4	10	1997, Ohio State	ADJB87	(A71)
CCGGG^{5br}CCCGG	$P2_12_12_1$	4	10	1997, Ohio State	ADJB80	(A71)
CCGGGCC^{5me}CGG	$P2_12_12_1$	4	10	1997, Ohio State	ADJB84,5	(A71)
C^{5me}CGGGCCCGG	$P6_1$	6	10	1997, Ohio State	ADJB86	(A71)
CCGGGCC^{5br}CGG	$P6_1$	6	10	1997, Ohio State	ADJB79	(A71)
CCGGGCC^{5me}CGG	$P6_1$	6	10	1997, Ohio State	ADJB83	(A71)

(*c*) Nonamers

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
GGATGGGAG	$P4_3$	4	9	1986, Cambridge	ADI009	(A14)

(*d*) Octamers, space group $P4_32_12$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
CCCCGGGG	8	4	1987, Weizmann/MIT	ADH012	(A16)
CCCCGGGG, 298 K	8	4	1995, Weizmann	ADH056	(A54)
CCCGCGGG	8	4	1997, Moscow	ADH0106	(A69)
CCCTAGGG	8	4	1996, Ohio State	ADH078	(A64)
GCCCGGGC	8	4	1987, Berlin	ADH008	(A17)
GCCC*GGGC (*methylenephosphonate)	8	4	1991, Berlin	ADHP36	(A36)
GGCCGGCC	8	4	1982, MIT	ADH013,098	(A4,5)
GGCCGGCC, 288 K	8	4	1995, Weizmann	ADH058	(A54)
GG^{5me}CCGGCC	8	4	1987, MIT	(ADHB21)	(A15)
GGGCGCCC, 293 K	8	4	1988, Weizmann	ADH026	(A22, A34)
GGGCGCCC, 115 K	8	4	1988, Weizmann	ADH027	(A20, A34)
GGGCGCCC, 115 K, re-refinement	8	4	1995, Weizmann	ADH057	(A54)
GTGCGCAC	8	4	1992, Ohio State	ADH047	(A40)
GTGTACAC/spermine	8	4	1987, Wisconsin	ADH014	(A18, A29)
CTCTAGAG	8	4	1989, Cambridge	ADH020	(A27)
GTACGTAC	8	4	1990, Kansas	ADH024	(A35)
GTACGTAC	8	4	1990, Bordeaux	ADH023	(A32)
GTCTAGAC	8	4	1992, Manchester	ADH041	(A42)
ATGCGCAT	8	4	1990, Institute of Cancer Research	(ADH032)	(A31)
ATGCGCAT/spermine	8	4	1990, Institute of Cancer Research	ADH033	(A31)
ACGTACGT	8	4	1996, Trinity, Dublin	ADH070	(A66)

Table A23.4.1.1 (continued)

(*e*) Octamers, space group $P2_12_12_1$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
CCCGCGGG	4	8	1997, Moscow	ADH0102–5	(A69)

(*f*) Octamers, space group $P6_1$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
GGGGCCCC	6	8	1985, Cambridge	ADH006	(A11)
GGGATCCC	6	8	1988, Berlin	ADH007	(A21)
GGGCGCCC, 293 K	6	8	1989, Weizmann	(ADH028)	(A30, A34)
GGGCGCCC, 100 K	6	8	1989, Weizmann	ADH029	(A30, A34)
GGGTACCC, 293 K	6	8	1990, Weizmann	ADH030	(A33)
GGGTACCC, 100 K	6	8	1990, Weizmann	ADH031	(A33)
GGGTGCCC	6	8	1988, Weizmann	ADH016	(A22)
GGTATACC	6	8	1981, Weizmann/Cambridge	ADH010	(A2, A7)
GG^{5br}UA^{5br}UACC	6	8	1981, Weizmann/Cambridge	ADHB11	(A2, A7, A13)
GGCATGCC	6	8	1997, Institute of Cancer Research	ADH076	(A70)
GGIGCTCC	6	8	1989, Cambridge	ADHB17	(A24)
GGGGCTCC mismatch	6	8	1985, Cambridge/Weizmann	ADH019	(A9, A12)
GGGGTCCC mismatch	6	8	1985, Cambridge/Weizmann	ADH018	(A10)
GGGTGCCC mismatch	6	8	1988, Weizmann	ADH016	(A22)

(*g*) Octamers, space group $P6_122$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
GTGTACAC	12	4	1989, Wisconsin	ADH034	(A28)
GTGTACAC/spermine	12	4	1993, Ohio State	ADH038	(A48)
GTGTACAC/spermidine	12	4	1993, Ohio State	ADH039	(A48)

(*h*) Octamers, space group $P2_12_12$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
GTACGTAC	4	8	1993, Bordeaux	ADH059	(A44)

(*i*) Hexamers, space group $C222_1$

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
GCCGGC	8	6	1995, Oregon State	ADF073	(A56)
G^{5me}CG^{5me}CGC	8	6	1995, Oregon State	ADFB62	(A56)
G^{5me}CCGGC	8	6	1995, Oregon State	ADFB63	(A56)
G^{5me}CGCGC	8	6	1995, Oregon State	ADFB72	(A56)

(*j*) Tetramers

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
^{5i}CCGG	$P4_32_12$	8	4	1981, UCLA (CIT)	ADDB01	(A1, A3, A8)

(*k*) RNA/DNA and RNA/RNA (lower case = RNA)

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
CCGGC g CCGG	$P2_12_12_1$	4	10	1994, Ohio State	AHJ052	(A49)
c CGGCGCCGg	$P2_12_12_1$	4	10	1994, Ohio State	AHJ060	(A50)
g CGTATACGC	$P2_12_12_1$	4	10	1993, MIT	AHJ043	(A45)
GCGTaTACGC	$P2_12_12_1$	4	10	1993, MIT	AHJ044	(A45)
GCGT^{me}aTACGC	$P2_12_12_1$	4	10	1994, ETH Zürich	AHJS55	(A53)
g c GTATACGC	$P2_12_12_1$	4	10	1995, MIT	AHJ068	(A55)
g c g TATACGC	$P2_12_12_1$	4	10	1982, MIT	AHJ015	(A4, A6)
g c g TATACCC\ \GGGTATACGC	$P2_12_12_1$	4	10	1992, MIT	AHJ040	(A43)
u u c g g g c g c c\ \GGCGCCCGAA	$P4_322$	8	10	1996, Upjohn	UHJ055	(A62)
c c c c g g g g	$P6_122$	12	4	1995, ETH Zürich	ARH063	(A57)
c c c c g g g g	$R32$	18	8	1995, ETH Zürich	ARH064	(A57)
c c c c g g g g	$R32$	18	8	1996, Northwestern	ARH074	(A61)
g u a u a u a C	$R3$	9	8	1996, Ohio State	AHH071	(A65)
g u a u g u a C	$R3$	9	8	1997, Ohio State	AHH077	(A68)
g u g u g u a C	$R3$	9	8	1997, Ohio State	AHH089	(A67)
g c u u c g g c ^{br}U	$C2$	4	9	1994, Cambridge	AHIB53	(A51)
(P)g g a c u u c g g u c c	$C2$	4	6	1991, Berkeley	ARL037	(A37)

Table A23.4.1.1 (continued)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
cgcgaattagcg	$P2_1$	2	12	1994, Manchester	ARL048	(A52)
uaaggaggugau	$P1$	1	24	1995, Berlin	ARL062	(A59)
ggcgcuugcguc	$P1$	1	24	1996, Colorado	URL050	(A63)
uuauauauauauaa	$P2_12_12_1$	4	4	1988, Strasbourg	ARN035	(A19, A25)

References (numbered chronologically by year and alphabetically by first author within each year)

Year	Reference
1981	(A1) R. E. Dickerson, H. R. Drew & B. N. Conner (1981). *Biomolecular stereodynamics*, Vol. 1, edited by R. H. Sarma, pp. 1–34. New York: Adenine Press.
	(A2) Z. Shakked, D. Rabinovich, W. B. T. Cruse, E. Egert, O. Kennard, G. Sala, S. A. Salisbury & M. A. Viswamitra (1981). *Proc. R. Soc. London Ser. B*, **213**, 479–487.
1982	(A3) B. N. Conner, T. Takano, S. Tanaka, K. Itakura & R. E. Dickerson (1982). *Nature (London)*, **295**, 294–299.
	(A4) S. Fujii, A. H.-J. Wang, J. van Boom & A. Rich (1982). *Nucleic Acids Res. Symp. Ser.* **11**, 109–112.
	(A5) A. H.-J. Wang, S. Fujii, J. H. van Boom & A. Rich (1982). *Proc. Natl Acad. Sci. USA*, **79**, 3968–3972.
	(A6) A. H.-J. Wang, S. Fujii, J. H. van Boom, G. A. van der Marel, S. A. A. van Boeckel & A. Rich (1982). *Nature (London)*, **299**, 601–604.
1983	(A7) Z. Shakked, D. Rabinovich, O. Kennard, W. B. T. Cruse, S. A. Salisbury & M. A. Viswamitra (1983). *J. Mol. Biol.* **166**, 183–201.
1984	(A8) B. N. Conner, C. Yoon, J. L. Dickerson & R. E. Dickerson (1984). *J. Mol. Biol.* **174**, 663–695.
1985	(A9) T. Brown, O. Kennard, G. Kneale & D. Rabinovich (1985). *Nature (London)*, **315**, 604–606.
	(A10) G. Kneale, T. Brown, O. Kennard & D. Rabinovich (1985). *J. Mol. Biol.* **186**, 805–814.
	(A11) M. McCall, T. Brown & O. Kennard (1985). *J. Mol. Biol.* **183**, 385–396.
1986	(A12) W. N. Hunter, G. Kneale, T. Brown, D. Rabinovich & O. Kennard (1986). *J. Mol. Biol.* **190**, 605–618.
	(A13) O. Kennard, W. B. T. Cruse, J. Nachman, T. Prange, Z. Shakked & D. Rabinovich (1986). *J. Biomol. Struct. Dyn.* **3**, 623–647.
	(A14) M. McCall, T. Brown, W. N. Hunter & O. Kennard (1986). *Nature (London)*, **322**, 661–664.
1987	(A15) C. A. Frederick, D. Saal, G. A. van der Marel, J. H. van Boom, A. H.-J. Wang & A. Rich (1987). *Biopolymers*, **26**, S145–S160.
	(A16) T. E. Haran, Z. Shakked, A. H.-J. Wang & A. Rich (1987). *J. Biomol. Struct. Dyn.* **5**, 199–217.
	(A17) U. Heinemann, H. Lauble, R. Frank & H. Bloeker (1987). *Nucleic Acids Res.* **15**, 9531–9550.
	(A18) S. Jain, G. Zon & M. Sundaralingam (1987). *J. Mol. Biol.* **197**, 141–145.
1988	(A19) A. C. Dock-Bregeon, B. Chevrier, A. Podjarny, D. Moras, J. S. de Bear, G. R. Gough, P. T. Gilham & J. E. Johnson (1988). *Nature (London)*, **335**, 375–378.
	(A20) M. Eisenstein, H. Hope, T. E. Haran, F. Frolow, Z. Shakked & D. Rabinovich (1988). *Acta Cryst.* B**44**, 625–628.
	(A21) H. Lauble, R. Frank, H. Bloecker & U. Heinemann (1988). *Nucleic Acids Res.* **16**, 7799–7816.
	(A22) D. Rabinovich, T. Haran, M. Eisenstein & Z. Shakked (1988). *J. Mol. Biol.* **200**, 151–161.
1989	(A23) C. A. Bingman, S. Jain, D. Jebaratnam & M. Sundaralingam (1989). Sixth Conversation in Biomolecular Stereodynamics, Albany, NY, Abstracts p. 28.
	(A24) W. B. T. Cruse, J. Aymani, O. Kennard, T. Brown, A. G. C. Jack & G. A. Leonard (1989). *Nucleic Acids Res.* **17**, 55–72.
	(A25) A. C. Dock-Bregeon, B. Chevrier, A. Podjarny, J. Johnson, J. S. de Bear, G. R. Gough, P. T. Gilham & D. Moras. (1989). *J. Mol. Biol.* **209**, 459–474.
	(A26) C. A. Frederick, G. J. Quigley, M.-K. Teng, M. Coll, G. A. van der Marel, J. H. van Boom, A. Rich & A. H.-J. Wang (1989). *Eur. J. Biochem.* **181**, 295–307.
	(A27) W. N. Hunter, B. L. D'Estaintot & O. Kennard (1989). *Biochemisty*, **28**, 2444–2451.
	(A28) S. Jain & M. Sundaralingam (1989). *J. Biol. Chem.* **264**, 12780–12784.
	(A29) S. Jain, G. Zon & M. Sundaralingam (1989). *Biochemistry*, **28**, 2360–2364.
	(A30) Z. Shakked, G. Guerstein-Guzikevich, F. Frolow & D. Rabinovich (1989). *Nature (London)*, **342**, 456–460.
1990	(A31) G. R. Clark, D. G. Brown, M. R. Sanderson, T. Chwalinski, S. Neidle, J. M. Veal, R. L. Jones, W. D. Wilson, G. Zon, E. Garman & D. I. Stuart (1990). *Nucleic Acids Res.* **18**, 5521–5528.
	(A32) C. Courseille, A. Dautant, M. Hospital, B. Langlois d'Estaintot, G. Precigoux, D. Molko & R. Teoule (1990). *Acta Cryst.* A**46**, FC9–FC12.
	(A33) M. Eisenstein, F. Frolow, Z. Shakked & D. Rabinovich (1990). *Nucleic Acids Res.* **18**, 3185–3194.
	(A34) Z. Shakked, G. Guerstein-Guzikevich, A. Zaytzev, M. Eisenstein, F. Frolow & D. Rabinovich (1990). In *Structure and methods*, Vol. 3. *DNA and RNA*, edited by R. H. Sarma & M. H. Sarma, pp. 55–72. Schenectady, NY: Adenine Press.
	(A35) F. Takusagawa (1990). *J. Biomol. Struct. Dyn.* **7**, 795–809.
1991	(A36) U. Heinemann, L.-N. Rudolph, C. Alings, M. Morr, W. Heikens, R. Frank & H. Bloecker (1991). *Nucleic Acids Res.* **19**, 427–433.
	(A37) S. R. Holbrook, C. Cheong, I. Tinoco Jr & S.-H. Kim (1991). *Nature (London)*, **353**, 579–581.
	(A38) N. Verdaguer, J. Aymami, D. Fernandez-Forner, I. Fita, M. Coll, T. Huynh-Dinh, J. Igolen & J. A. Subirana (1991). *J. Mol. Biol.* **221**, 623–635.
1992	(A39) C. Bingman, S. Jain, G. Zon & M. Sundaralingam (1992). *Nucleic Acids Res.* **20**, 6637–6647.
	(A40) C. A. Bingman, X. Li, G. Zon & M. Sundaralingam (1992). *Biochemistry*, **31**, 12803–12812.
	(A41) C. A. Bingman, G. Zon & M. Sundaralingam (1992). *J. Mol. Biol.* **227**, 738–756.
	(A42) A. R. Cervi, B. Langlois d'Estaintot & W. N. Hunter (1992). *Acta Cryst.* B**48**, 714–719.
	(A43) M. Egli, N. Usman, S. Zhang & A. Rich (1992). *Proc. Natl Acad. Sci. USA*, **89**, 534–538.
1993	(A44) B. Langlois D'Estaintot, A. Dautant, C. Courseille & G. Precigoux (1993). *Eur. J. Biochem.* **213**, 673–682.
	(A45) M. Egli, N. Usman & A. Rich (1993). *Biochemistry*, **32**, 3221–3237.
	(A46) B. Ramakrishnan & M. Sundaralingam (1993). *Biochemistry*, **32**, 11458–11468.
	(A47) B. Ramakrishnan & M. Sundaralingam (1993). *J. Mol. Biol.* **231**, 431–444.
	(A48) N. Thota, X. H. Li, C. Bingman & M. Sundaralingam (1993). *Acta Cryst.* D**49**, 282–291.
1994	(A49) C. Ban, B. Ramakrishnan & M. Sundaralingam (1994). *J. Mol. Biol.* **236**, 275–285.
	(A50) C. Ban, B. Ramakrishnan & M. Sundaralingam (1994). *Nucleic Acids Res.* **22**, 5466–5476.
	(A51) W. Cruse, P. Saludjian, E. Biala, P. Strazewski, T. Prange & O. Kennard (1994). *Proc. Natl Acad. Sci. USA*, **91**, 4160–4164.
	(A52) G. A. Leonard, K. E. McAuley-Hecht, S. Ebel, D. M. Lough, T. Brown & W. N. Hunter (1994). *Structure*, **2**, 483–494.
	(A53) P. Lubini, W. Zuercher & M. Egli (1994). *Chem. Biol.* **1**, 39–45.

Table A23.4.1.1 (continued)

Year	Reference
1995	(A54) M. Eisenstein & Z. Shakked (1995). *J. Mol. Biol.* **248**, 662–678.
	(A55) Y.-G. Gao, H. Robinson, J. H. van Boom & A. H.-J. Wang (1995). *Biophys. J.* **69**, 559–568.
	(A56) B. H. Mooers, G. P. Schroth, W. W. Baxter & P. S. Ho (1995). *J. Mol. Biol.* **249**, 772–784.
	(A57) S. Portmann, N. Usman & M. Egli (1995). *Biochemistry*, **34**, 7569–7575.
	(A58) B. Ramakrishnan & M. Sundaralingam (1995). *Biophys. J.* **69**, 553–558.
	(A59) H. Schindelin, M. Zhang, R. Bald, J.-P. Fuerste, V. A. Erdmann & U. Heinemann (1995). *J. Mol. Biol.* **249**, 595–603.
1996	(A60) C. Ban & M. Sundaralingam (1996). *Biophys. J.* **71**, 1222–1227.
	(A61) M. Egli, S. Portmann & N. Usman (1996). *Biochemistry*, **35**, 8489–8494.
	(A62) N. C. Horton & B. C. Finzel (1996). *J. Mol. Biol.* **264**, 521–533.
	(A63) S. E. Lietzke, C. L. Barnes & C. E. Kundrot (1996). *Structure*, **4**, 917–930.
	(A64) D. B. Tippin & M. Sundaralingam (1996). *Acta Cryst.* D**52**, 997–1003.
	(A65) M. C. Wahl, C. Ban, C. Sekharudu, B. Ramakrishnan & M. Sundaralingam (1996). *Acta Cryst.* D**52**, 655–667.
	(A66) D. J. Wilcock, A. Adams, C. J. Cardin & L. P. G. Wakelin (1996). *Acta Cryst.* D**52**, 481–485.
1997	(A67) R. Biswas & M. Sundaralingam (1997). *J. Mol. Biol.* **270**, 511–519.
	(A68) R. Biswas, M. C. Wahl, C. Ban & M. Sundaralingam (1997). *J. Mol. Biol.* **267**, 1149–1156.
	(A69) L. G. Fernandez, J. A. Subirana, N. Verdagauer, D. Pyshni, L. Campos & L. Malinina (1997). *J. Biomol. Struct. Dyn.* **15**, 151–163.
	(A70) C. M. Nunn & S. Neidle (1997). *Acta Cryst.* D**53**, 269–273.
	(A71) D. B. Tippin & M. Sundaralingam (1997). *J. Mol. Biol.* **267**, 1171–1185.

Table A23.4.1.2

X-ray analyses of B-DNA helices and their complexes with minor-groove-binding drug molecules

See introductory notes to Table A23.4.1.1. Space group $P2_12_12_1$ unless specified otherwise.
Notes: (triplet) = external triplet formed from overhanging bases. Overhanging, unpaired bases are $\underline{\underline{\text{double underlined}}}$. $\underline{\text{Single underlining}}$ calls attention to interesting or relevant sequence aspects. Other notes as in Table A23.4.1.1.

I. DNA duplexes without bound drugs

(*a*) Dodecamers, space group $P2_12_12_1$

(1) Oligonucleotides without mismatches

Sequence	*Z*	Ubp	Date, institution	NDB No.	Reference
$\text{CGCG}\underline{\text{AATT}}\text{CGCG}$, 290 K	4	12	1980, UCLA (CIT)	BDL001	(B1–5, B75)
$\text{CGCG}\underline{\text{AATT}}\text{CGCG}$, 16 K	4	12	1982, UCLA (CIT)	BDL002	(B6)
$\text{CGCG}\underline{\text{AATT}}\text{CGCG}$, re-refinement	4	12	1987, Strasbourg	BDL020	(B23)
$\text{CGCG}\underline{\text{AATT}}\text{CGCG}$, anisotropic temperature-factor refinement	4	12	1985, Berkeley	BDL005	(B10)
$\text{CGCG}\underline{\text{AATT}}^{\text{5br}}\text{CGCG}$, 293 K	4	12	1982, UCLA (CIT)	BDLB03	(B7, B8)
$\text{CGCG}\underline{\text{AATT}}^{\text{5br}}\text{CGCG}$, 280 K	4	12	1982, UCLA (CIT)	BDLB04	(B7, B8, B75)
$\text{CGCG}\underline{\text{A}}^{\text{6me}}\underline{\text{ATT}}\text{CGCG}$	4	12	1988, MIT	BDLB13	(B24)
$\text{CGCG}\underline{\text{AA}}^{\text{6ame}}\underline{\text{T}}^{\text{6ame}}\underline{\text{T}}\text{CGCG}$	4	12	1997, Northwestern	BDLS79	(B111)
$\text{CGCG}\underline{\text{AA}}^{\text{6aOH}}\underline{\text{T}}^{\text{6aOH}}\underline{\text{T}}\text{CGCG}$	4	12	1997, Northwestern	BDLS80	(B111)
$\text{CGCG}\underline{\text{AASS}}\text{CGCG}$	4	12	1996, Manchester	BDLS67	(B97)
$\text{CGCAIAT}^{\text{5me}}\text{CTGCG}$	4	12	1997, Weizmann	BDLB82	(B113)
$\text{CGC}\underline{\text{AAAAAA}}\text{GCG}$	4	12	1987, Cambridge	BDL006	(B20, B75)
$\text{CGC}\underline{\text{AAAAAT}}\text{GCG}$	4	12	1989, Yale	BDL015	(B31, B75)
$\text{CGC}\underline{\text{AAATTT}}\text{GCG}$	4	12	1987, MIT	BDL016	(B17)
$\text{CGC}\underline{\text{AAATTT}}\text{GCG}$	4	12	1992, Institute of Cancer Research	BDL038	(B52, B75)
$\text{CGC}\underline{\text{ATATAT}}\text{GCG}$	4	12	1988, UCLA	BDL007	(B27)
$\text{CGCG}\underline{\text{TTAA}}\text{CGCG}$	4	12	1991, Ohio State	BDL059	(B40, B86)
$\text{CGCG}\underline{\text{ATAT}}\text{CGCG}$	4	12	1997, Weizmann	BDL078	(B113)
$\text{CGCAIAT}^{\text{5me}}\text{CTGCG}$	4	12	1997, Weizmann	BDLB76	(B113)
$\text{CG}\underline{\text{T}}\text{G}\underline{\text{AATT}}\text{C}\underline{\text{A}}\text{CG}$	4	12	1991, UCLA	BDL029	(B44, B75)
$\text{CG}\underline{\text{T}}\text{G}\underline{\text{AATT}}\text{C}\underline{\text{A}}\text{CG}$	4	12	1991, Rutgers	BDL028	(B45)
$\text{CGCG}\underline{\text{AAAA}}\text{CGCG/ CGCG}\underline{\text{TT/TT}}\text{CGCG}$ (nicked strand)	4	12	1990, MIT	BDL021,32	(B35)

(2) Mismatch oligonucleotides (mismatches underlined)

Sequence	*Z*	Ubp	Date, institution	NDB No.	Reference
$\text{CGC}\underline{\text{G}}\text{AATT}\underline{\text{G}}\text{GCG}$	4	12	1993, Institute of Cancer Research	BDL046	(B72)
$\text{CGC}\underline{\text{G}}\text{AATT}\underline{\text{A}}\text{GCG}$	4	12	1986, Cambridge	BDL012	(B13, B15)
$\text{CGC}\underline{\text{G}}\text{AATT}^{\text{6et}}\underline{\text{A}}\text{GCG}$	4	12	1994, Manchester	BDLB54	(B79)
$\text{CGC}\underline{\text{G}}\text{AATT}^{\text{8oxo}}\underline{\text{A}}\text{GCG}$	4	12	1992, Manchester	BDLB33	(B57)
$\text{CGC}\underline{\text{G}}\text{AATT}\underline{\text{T}}\text{GCG}$	4	12	1985, Cambridge	BDL009	(B19)
$\text{CGC}^{\text{6me}}\underline{\text{G}}\text{AATT}\underline{\text{T}}\text{GCG}$	4	12	1990, Edinburgh	BDLB26	(B38)
$\text{CGC}\underline{\text{A}}\text{ATT}\underline{\text{G}}\text{GCG}$	4	12	1989, Manchester	BDL014	(B28, B37)
$\text{CGC}\underline{\text{A}}\text{AGCT}\underline{\text{G}}\text{GCG}$	4	12	1990, Institute of Cancer Research	BDL022	(B39, B75)
$\text{CGC}\underline{\text{A}}\text{AATT}^{\text{8oxo}}\underline{\text{G}}\text{GCG}$	4	12	1994, Edinburgh	BDLB56	(B80)
$\text{CGC}\underline{\text{A}}\text{AATT}\underline{\text{C}}\text{GCG}$	4	12	1986, Cambridge	BDL011	(B16)

Table A23.4.1.2 (continued)

Sequence	Z	Ubp	Date, institution	NDB No.	Reference
CGC$\underline{\text{A}}$AATT$\underline{\text{I}}$GCG	4	12	1992, Edinburgh	BDLB41	(B56)
CGC$\underline{\text{I}}$AATT$\underline{\text{A}}$GCG	4	12	1987, Cambridge	BDLB10	(B18)
CGC$\underline{\text{I}}$AATT$\underline{\text{C}}$GCG	4	12	1992, Thomas Jefferson	BDLB40	(B61)
CG$\underline{\text{A}}$GAATTC$^{\text{6me}}\underline{\text{G}}$CG	4	12	1994, Rutgers	BDLB53	(B76)
CG$\underline{\text{T}}$GAATTC$^{\text{6me}}\underline{\text{G}}$CG	4	12	1995, Rutgers	BDLB58	(B95)

(*b*) Dodecamers: other space groups

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
CGCTCTAGAGCG	$P2_1$	2	24	1996, Barcelona	BDL070	(B102)
CGTAGATCTACG	$C2$	4	12	1993, Manchester	BDL042	(B69, B75)
CGCG$\underline{\text{AAAAAA}}$CG	$P2_12_12$	4	24	1993, Yale	BDL047	(B64, B75)
ACCGGCGCCACA	$R3$	9	12	1989, Strasbourg	BDL018	(B34, B48, B49)
ACCGCCGGCGCC	$R3$	9	12	1989, Strasbourg	BDL035	(B48, B49)
ACCGC$^{\text{5me}}$CGGCGCC	$R3$	9	12	1997, Strasbourg	BDLB83	(B109)
ACCGGCGCCACA	$R3$	9	12	1991, Strasbourg	BDL034	(B48)

(*c*) Decamers

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
CCAA$\underline{\text{GA}}$TTGG mismatch	$C2$	4	5	1987, UCLA	BDJ008	(B22, B25)
CCAA$\underline{\text{CG}}$TTGG, Mg	$C2$	4	5	1991, UCLA	BDJ019	(B46, B50)
CCAA$\underline{\text{CI}}$TTGG, Ca	$C2$	4	5	1992, UCLA	BDJB44	(B70)
CCA$\underline{\text{GGCC}}$TGG	$C2$	4	5	1989, Berlin	BDJ017	(B32)
CCA$\underline{\text{GGC}}^{\text{ara}}\underline{\text{C}}$TGG	$C2$	4	5	1991, MIT	BDJS30	(B41)
CCA$^{\text{8oxo}}\underline{\text{GCGC}}$TGG	$C2$	4	5	1995, MIT	BDJB57	(B91)
CTCTCGAGAG	$C2$	4	10	1994, UCLA	BDJ060	(B89)
CGC$\underline{\text{AATT}}$GCG	$C2$	4	10	1997, Institute of Cancer Research	BDJ069	(B114)
C$\underline{\text{AAA}}$G$\underline{\text{AAAA}}$G	$C2$	4	20	1997, UCLA	BDJ081	(B107)
$\underline{\underline{\text{CG}}}$ACGATCGT TGCTAGCA$\underline{\underline{\text{GC}}}$	$P2_1$	2	10	1997, NYU	UDJ060	(B112)
$\underline{\underline{\text{GG}}}CC\underline{\text{AATT}}$GG GG$\underline{\text{TTAA}}CC\underline{\underline{\text{GG}}}$	$P2_12_12_1$	4	10	1996, Cambridge	UDJ049	(B103)
CG$\underline{\text{ATCGAT}}$CG, Mg	$P2_12_12_1$	4	10	1991, UCLA	BDJ025	(B42)
CG$\underline{\text{ATTAAT}}$CG, Mg	$P2_12_12_1$	4	10	1992, UCLA	BDJ031	(B58)
CG$\underline{\text{ATATAT}}$CG, Mg	$P2_12_12_1$	4	10	1992, UCLA	BDJ037	(B62)
CG$\underline{\text{ATATAT}}$CG, Ca	$P2_12_12_1$	4	10	1992, UCLA	BDJ036	(B62)
CAT$\underline{\text{GGCC}}$ATG, Ca	$P2_12_12_1$	4	10	1993, UCLA	BDJ051	(B66)
CG$\underline{\text{ATCG}}^{\text{6me}}\underline{\text{AT}}$CG	$P3_221$	6	10	1992, UCLA	BDJB48	(B63)
CCAA$\underline{\text{CI}}$TTGG, Mg	$P3_221$	6	10	1992, UCLA	BDJB43	(B70)
CC$\underline{\text{ATTAAT}}$GG, Mg	$P3_221$	6	10	1994, UCLA	BDJ055	(B77)
CCACTAGTGG	$P3_221$	6	10	1994, Weizmann	BDJ061	(B82)
CCAGGC$^{\text{5me}}\underline{\text{C}}$TGG	$P6$	6	10	1992, Berlin	BDJB27	(B43, B54)
CCA$\underline{\text{GGC}}^{\text{5me}}\underline{\text{C}}$TGG	$P6$	6	10	1993, Berlin	BDJB49	(B68)
CCA$\underline{\text{GGC}}^{\text{5me}}\underline{\text{C}}$TGG	$P6$	6	10	1993, Berlin	BDJB50	(B68)
CCAA$\underline{\text{GC}}$TTGG	$P6$	6	10	1993, UCLA	BDJ052	(B67)
CCGGCGCCGG	$R3$	9	10	1992, Berlin	BDJ039	(B55)
CCGCCGGCGG	$R3$	9	10	1994, Strasbourg	BD0015	(B85)
CCIIICCCGG	$P3_1$	3	10	1997, Weizmann	BDJB77	(B113)

(*d*) Other oligonucleotide lengths

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
$\underline{\underline{\text{G}}}CG\underline{\text{AATT}}$CG (triplet) GC$\underline{\text{TTAA}}GC\underline{\underline{\text{G}}}$	$P2_12_12_1$	4	8	1996, Cambridge	UDI030	(B94)
CGCTAGCG	$P2_12_12_1$	4	16	1996, Barcelona	BDH071	(B102)
$\underline{\underline{\text{C}}}GG\underline{\text{T}}$GG CC$\underline{\text{A}}CC\underline{\underline{\text{G}}}$	$P6_122$	12	6	1995, Manitoba	BDF062	(B93)
CTCGAG	$P6_222$	12	3	1996, Ohio State	BDF068	(B104)
GpsCGpsCGpsC	$P2_12_12_1$	4	6	1987, Cambridge	BDFP24	(B14)

II. DNA complexes with minor-groove-binding drugs

(*a*) Netropsin family of polyamides

Sequence	Space group	Z	Ubp	Date, institution	NDB No.	Reference
Netropsin: $^{+}$Py-Py^{+}						
CGCG$\underline{\text{AATT}}^{\text{5br}}$CGCG/N		4	12	1985, UCLA	GDLB05	(B11, B12)
CGCG$\underline{\text{AATT}}^{\text{5br}}$CGCG/N		4	12	1995, UCLA	GDLB31	(B88)
CGCG$\underline{\text{AATT}}$CGCG/N		4	12	1992, Illinois	GDL018	(B59)
CGC$^{\text{6et}}$G$\underline{\text{AATT}}$CGCG/N		4	12	1992, Illinois	GDLB17	(B59)

Table A23.4.1.2 (continued)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCAAATTTGCG/N		4	12	1993, MIT	GDL014	(B73)
CGCGATATCGCG/N		4	12	1989, MIT	GDL001,4	(B30)
CGCGTTAACGCG/N		4	12	1995, Ohio State	GDL030	(B86)
CGCAATTGCG/N		4	12	1997, Institute of Cancer Research	GDJ046	(B110)
Lexitropsin: $^{+}$Im-Py^{+}						
CGCGAATTCGCG/1L		4	12	1995, UCLA	GDL037,8	(B90)
2:1 Di-imidazole Lexitropsin: ^{0}Im-Im^{+}						
CATGGCCATG/2D		4	10	1997, UCLA	GDJ054	(B108)
Distamycin: ^{0}Py-Py-Py^{+}						
CGCAAATTTGCG/1D		4	12	1987, MIT	GDL003	(B17)
ICICICIC/2D	$P4_122$	8	4	1994, Ohio State	GDHB25	(B74)
IcICICIC/2D	$P4_122$	8	4	1995, Ohio State	GHHB34	(B87)
IcIcICIC/2D	$P4_122$	8	4	1995, Ohio State	GHHB35	(B87)
ICATATIC	$P4_122$	8	4	1997, Ohio State	GHHB50	(B105)
ICITACIC	$P4_122$	8	4	1997, Ohio State	GHHB51	(B105)
ICATATIC	*C*2	4	4	1997, Ohio State	GDLB49	(B105)

(*b*) Hoechst family

Sequence	*Z*	Ubp	Date, institution	NDB No.	Reference
Hoechst 33258 (*para* -OH on phenyl ring A)					
CGCGAATTCGCG/H	4	12	1987, UCLA	GDL006	(B21)
CGCGAATTCGCG/H	4	12	1988, MIT	GDL002	(B26)
CGCGAATTCGCG/H, 273 K	4	12	1991, UCLA	GDL010,11	(B47)
CGCGAATTCGCG/H, 248 K	4	12	1991, UCLA	GDL012	(B47)
CGCGAATTCGCG/H, 173 K	4	12	1991, UCLA	GDL013	(B47)
CGCGATATCGCG/H	4	12	1989, MIT	GDL007	(B29)
CGCAAATTTGCG/H	4	12	1994, Institute of Cancer Research	GDL028	(B83)
CGCAAATTTGCG/H	4	12	1994, MIT	GDL026	(B84)
CGCGAATTCGCG/H	4	12	1992, Illinois	GDL022	(B60)
CGC^{6et}GAATTCGCG/H	4	12	1992, Illinois	GDLB19	(B60)
Meta-OH(N) Hoechst 33258 (*meta* -OH on ring A)					
CGCGAATTCGCG/H 'in'	4	12	1996, Institute of Cancer Research	GDL047	(B99)
CGCGAATTCGCG/H 'out'	4	12	1996, Institute of Cancer Research	GDL048	(B99)
Hoechst 33342 (*para* -OEt on ring A)					
CGCGAATTCGCG/H	4	12	1992, Illinois	GDLB20	(B60)
CGC^{6et}GAATTCGCG/H	4	12	1992, Illinois	GDLB20	(B60)
Bis-benzimidazole compound (imidazole for piperazine on Hoechst 33258)					
CGCGAATTCGCG/B	4	12	1995, Institute of Cancer Research	GDL033	(B96)
Tribiz or Tris-benzimidazole (extended Hoechst 33258 analogue)					
CGCAAATTTGCG/T	4	12	1996, Institute of Cancer Research	GDL039	(B98)
Bis-amidinium derivative of Hoechst 33258					
CGCGAATTCGCG	4	12	1997, Institute of Cancer Research	GDL052	(B106)

(*c*) Berenil family

Sequence	*Z*	Ubp	Date, institution	NDB No.	Reference
Berenil					
CGCGAATTCGCG/B	4	12	1990, Institute of Cancer Research	GDL009	(B36)
CGCGAATTCGCG/B	4	12	1992, Institute of Cancer Research	GDL016	(B51)
2,5-Bis(4-guanylphenyl)furan (berenil analogue)					
CGCGAATTCGCG/F	4	12	1996, Institute of Cancer Research	GDL036	(B100)
2,5-Bis{[4-(*N*-isopropyl)amidino]phenyl}furan (berenil analogue)					
CGCGAATTCGCG/F	4	12	1996, Institute of Cancer Research	GDL044	(B101)
2,5-Bis{[4-(*N*-cyclopropyl)amidino]phenyl}furan (berenil analogue)					
CGCGAATTCGCG/F	4	12	1997, Institute of Cancer Research	GDL045	(B101)

Table A23.4.1.2 (continued)

(*d*) Other minor-groove binders

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
DAPI						
CGCGAATTCGCG/D		4	12	1989, UCLA	GDL008	(B33)
Pentamidine						
CGCGAATTCGCG/P		4	12	1992, Institute of Cancer Research	GDL015	(B53)
γ-Oxapentamidine						
CGCGAATTCGCG/P		4	12	1994, Institute of Cancer Research	GDL027	(B81)
Propamidine						
CGCGAATTCGCG/P		4	12	1993, Institute of Cancer Research	GDL023	(B71)
CGCGAATTCGCG/P		4	12	1995, Institute of Cancer Research	GDL032	(B92)
SN6999						
CGC^{6et}GAATTCGCG/S		4	12	1993, Illinois	GDLB24	(B65)
Anthramycin						
CCAACGTTGG/A	$P3_221$	6	5	1993, UCLA	GDJB29	(B78)

References (numbered chronologically by year and alphabetically by first author within each year)

Year	Reference
1980	(B1) R. M. Wing, H. R. Drew, T. Takano, C. Broka, S. Tanaka, K. Itakura & R. E. Dickerson (1980). *Nature (London)*, **287**, 755–758.
1981	(B2) R. E. Dickerson & H. R. Drew (1981). *J. Mol. Biol.* **149**, 761–786.
	(B3) R. E. Dickerson, H. R. Drew & B. N. Conner (1981). *Biomolecular stereodynamics*, Vol. 1, edited by R. H. Sarma, pp. 1–34. New York: Adenine Press.
	(B4) H. R. Drew & R. E. Dickerson (1981). *J. Mol. Biol.* **151**, 535–556.
	(B5) H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura & R. E. Dickerson (1981). *Proc. Natl Acad. Sci. USA*, **78**, 2179–2183.
1982	(B6) H. R. Drew, S. Samson & R. E. Dickerson (1982). *Proc. Natl Acad. Sci. USA*, **79**, 4040–4044.
	(B7) A. V. Fratini, M. L. Kopka, H. R. Drew & R. E. Dickerson (1982). *J. Biol. Chem.* **257**, 14686–14707.
1983	(B8) M. L. Kopka, A. V. Fratini, H. R. Drew & R. E. Dickerson (1983). *J. Mol. Biol.* **163**, 129–146.
1984	(B9) R. M. Wing, P. Pjura, H. R. Drew & R. E. Dickerson (1984). *EMBO J.* **3**, 1201–1206.
1985	(B10) S. R. Holbrook, R. E. Dickerson & S.-H. Kim (1985). *Acta Cryst.* B**41**, 255–262.
	(B11) M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura & R. E. Dickerson (1985). *Proc. Natl Acad. Sci. USA*, **82**, 1376–1380.
	(B12) M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura & R. Dickerson (1985). *J. Mol. Biol.* **183**, 553–563.
1986	(B13) T. Brown, W. N. Hunter, G. Kneale & O. Kennard (1986). *Proc. Natl Acad. Sci. USA*, **83**, 2402–2406.
	(B14) W. B. T. Cruse, S. A. Salisbury, T. Brown, R. Cosstick, F. Eckstein & O. Kennard (1986). *J. Mol. Biol.* **192**, 891–905.
	(B15) W. N. Hunter, T. Brown & O. Kennard (1986). *J. Biomol. Struct. Dyn.* **4**, 173–191.
	(B16) W. N. Hunter, T. Brown, N. N. Anand & O. Kennard (1986). *Nature (London)*, **320**, 552–555.
1987	(B17) M. Coll, C. A. Frederick, A. H.-J. Wang & A. Rich (1987). *Proc. Natl Acad. Sci. USA*, **84**, 8385–8389.
	(B18) P. W. R. Corfield, W. N. Hunter, T. Brown, P. Robinson & O. Kennard (1987). *Nucleic Acids Res.* **15**, 7935–7949.
	(B19) W. N. Hunter, T. Brown, G. Kneale, N. N. Anand, D. Rabinovich & O. Kennard (1987). *J. Biol. Chem.* **21**, 9962–9970.
	(B20) H. C. M. Nelson, J. T. Finch, B. F. Luisi & A. Klug (1987). *Nature (London)*, **330**, 221–226.
	(B21) P. E. Pjura, K. Grzeskowiak & R. E. Dickerson (1987). *J. Mol. Biol.* **197**, 257–271.
	(B22) G. G. Privé, U. Heinemann, S. Chandrasegaran, L.-S. Kan, M. L. Kopka & R. E. Dickerson (1987). *Science*, **238**, 498–504.
	(B23) E. Westhof (1987). *J. Biomol. Struct. Dyn.* **5**, 581–600.
1988	(B24) C. A. Frederick, G. J. Quigley, G. A. van der Marel, J. H. van Boom, A. H.-J. Wang & A. Rich (1988). *J. Biol. Chem.* **263**, 17872–17879.
	(B25) G. G. Privé, U. Heinemann, S. Chandrasegaran, L.-S. Kan, M. L. Kopka & R. E. Dickerson (1988). *Structure and expression*, Vol. 2. *DNA and its drug complexes*, edited by R. H. Sarma & M. H. Sarma, pp. 27–47. Schenectady, NY: Adenine Press.
	(B26) M. Teng, N. Usman, C. A. Frederick & A. H.-J. Wang (1988). *Nucleic Acids Res.* **16**, 2671–2690.
	(B27) C. Yoon, G. G. Privé, D. S. Goodsell & R. E. Dickerson (1988). *Proc. Natl Acad. Sci. USA*, **85**, 6332–6336.
1989	(B28) T. Brown, G. A. Leonard, E. D. Booth & J. Chambers (1989). *J. Mol. Biol.* **207**, 455–457.
	(B29) M. A. A. F. de C. T. Carrondo, M. Coll, J. Aymami, A. H.-J. Wang, G. A. van der Marel, J. H. van Boom & A. Rich (1989). *Biochemistry*, **28**, 7849–7859.
	(B30) M. Coll, J. Aymami, G. A. van der Marel, J. H. van Boom, A. Rich & A. H.-J. Wang (1989). *Biochemistry*, **28**, 310–320.
	(B31) A. D. DiGabriele, M. R. Sanderson & T. A. Steitz (1989). *Proc. Natl Acad. Sci. USA*, **86**, 1816–1820.
	(B32) U. Heinemann & C. Alings (1989). *J. Mol. Biol.* **210**, 369–381.
	(B33) T. A. Larsen, D. S. Goodsell, D. Cascio, K. Grzeskowiak & R. E. Dickerson (1989). *J. Biomol. Struct. Dyn.* **7**, 477–491.
	(B34) Y. Timsit, E. Westhof, R. P. P. Fuchs & D. Moras (1989). *Nature (London)*, **341**, 459–462.
1990	(B35) J. Aymami, M. Coll, G. A. van der Marel, J. H. van Boom, A. H.-J. Wang & A. Rich (1990). *Proc. Natl Acad. Sci. USA*, **87**, 2526–2530.
	(B36) D. G. Brown, M. R. Sanderson, J. V. Skelly, T. C. Jenkins, T. Brown, E. Garman, D. I. Stuart & S. Neidle (1990). *EMBO J.* **9**, 1329–1334.
	(B37) G. A. Leonard, E. D. Booth & T. Brown (1990). *Nucleic Acids Res.* **18**, 5617–5623.
	(B38) G. A. Leonard, J. Thomson, W. P. Watson & T. Brown (1990). *Proc. Natl Acad. Sci. USA*, **87**, 9573–9576.
	(B39) G. D. Webster, M. R. Sanderson, J. V. Skelly, S. Neidle, P. F. Swann, B. F. Li & I. J. Tickle (1990). *Proc. Natl Acad. Sci. USA*, **87**, 6693–6697.
1991	(B40) J. Balendrian & M. Sundaralingam (1991). *J. Biomol. Struct. Dyn.* **9**, 511–516.
	(B41) Y.-G. Gao, G. A. van der Marel, J. H. van Boom & A. H.-J. Wang (1991). *Biochemistry*, **30**, 9922–9931.
	(B42) K. Grzeskowiak, K. Yanagi, G. G. Privé & R. E. Dickerson (1991). *J. Biol. Chem.* **266**, 8861–8883.
	(B43) U. Heinemann & C. Alings (1991). *EMBO J.* **10**, 35–43.
	(B44) T. A. Larsen, M. L. Kopka & R. E. Dickerson (1991). *Biochemistry*, **30**, 4443–4449.
	(B45) N. Narayana, S. L. Ginell, I. M. Russu & H. M. Berman (1991). *Biochemistry*, **30**, 4450–4455.

Table A23.4.1.2 (continued)

Year	Reference
	(B46) G. G. Privé, K. Yanagi & R. E. Dickerson (1991). *J. Mol. Biol.* **217**, 177–199.
	(B47) J. R. Quintana, A. A. Lipanov & R. E. Dickerson (1991). *Biochemistry*, **30**, 10294–10306.
	(B48) Y. Timsit, E. Vilbois & D. Moras (1991). *Nature (London)*, **354**, 167–170.
	(B49) Y. Timsit & D. Moras (1991). *J. Mol. Biol.* **221**, 919–940.
	(B50) K. Yanagi, G. D. Privé & R. E. Dickerson (1991). *J. Mol. Biol.* **217**, 201–214.
1992	(B51) D. G. Brown, M. R. Sanderson, E. Garman & S. Neidle (1992). *J. Mol. Biol.* **226**, 481–490.
	(B52) K. J. Edwards, D. G. Brown, N. Spink, J. V. Skelly & S. Neidle (1992). *J. Mol. Biol.* **226**, 1161–1173.
	(B53) K. J. Edwards, T. C. Jenkins & S. Neidle (1992). *Biochemistry*, **31**, 7104–7109.
	(B54) U. Heinemann & M. Hahn (1992). *J. Biol. Chem.* **267**, 7332–7341.
	(B55) U. Heinemann, C. Alings & M. Bansal (1992). *EMBO J.* **11**, 1931–1939.
	(B56) G. A. Leonard, E. D. Booth, W. N. Hunter & T. Brown (1992). *Nucleic Acids Res.* **20**, 4753–4759.
	(B57) G. A. Leonard, A. Guy, T. Brown, R. Teoule & W. N. Hunter (1992). *Biochemistry*, **31**, 8415–8420.
	(B58) J. R. Quintana, K. Grzeskowiak, K. Yanagi & R. E. Dickerson (1992). *J. Mol. Biol.* **225**, 379–395.
	(B59) M. Sriram, G. A. van der Marel, H. L. P. F. Roelen, J. H. van Boom & A. H.-J. Wang (1992). *Biochemistry*, **31**, 11823–11834.
	(B60) M. Sriram, G. A. van der Marel, H. L. P. F. Roelen, J. H. van Boom & A. H.-J. Wang (1992). *EMBO J.* **11**, 225–232.
	(B61) J.-C. Xuan & I. T. Weber (1992). *Nucleic Acids Res.* **20**, 5457–5464.
	(B62) H. Yuan, J. R. Quintana & R. E. Dickerson (1992). *Biochemistry*, **31**, 8009–8021.
1993	(B63) I. Baikalov, K. Grzeskowiak, K. Yanagi, J. Quintana & R. E. Dickerson (1993). *J. Mol. Biol.* **231**, 768–784.
	(B64) A. D. DiGabriele & T. A. Steitz (1993). *J. Mol. Biol.* **231**, 1024–1029.
	(B65) Y.-G. Gao, M. Sriram, W. A. Denny & A. H.-J. Wang (1993). *Biochemistry*, **32**, 9693–9648.
	(B66) D. S. Goodsell, M. L. Kopka, D. Cascio & R. E. Dickerson (1993). *Proc. Natl Acad. Sci. USA*, **90**, 2930–2934.
	(B67) K. Grzeskowiak, D. S. Goodsell, M. Kaczor-Grzeskowiak, D. Cascio & R. E. Dickerson (1993). *Biochemistry*, **32**, 8923–8931.
	(B68) M. Hahn & U. Heinemann (1993). *Acta Cryst.* D**49**, 468–477.
	(B69) G. A. Leonard & W. N. Hunter (1993). *J. Mol. Biol.* **234**, 198–208.
	(B70) A. Lipanov, M. L. Kopka, M. Kaczor-Grzeskowiak, J. Quintana & R. E. Dickerson (1993). *Biochemistry*, **32**, 1373–1389.
	(B71) C. M. Nunn, T. C. Jenkins & S. Neidle (1993). *Biochemistry*, **32**, 13838–13842.
	(B72) J. V. Skelly, K. J. Edwards, T. C. Jenkins & S. Neidle (1993). *Proc. Natl Acad. Sci. USA*, **90**, 804–808.
	(B73) L. Tabernero, N. Verdaguer, M. Coll, I. Fita, G. A. van der Marel, J. H. van Boom, A. Rich & J. Aymami (1993). *Biochemistry*, **32**, 8403–8410.
1994	(B74) X. Chen, B. Ramakrishnan, S. T. Rao & M. Sundaralingam (1994). *Nature Struct. Biol.* **1**, 169–170.
	(B75) R. E. Dickerson, D. S. Goodsell & S. A. Neidle (1994). *Proc. Natl Acad. Sci. USA*, **91**, 3579–3583.
	(B76) S. L. Ginnell, J. Vojtechovsky, B. Gaffney, R. Jones & H. M. Berman (1994). *Biochemistry*, **33**, 3487–3493.
	(B77) D. S. Goodsell, M. Kaczor-Grzeskowiak & R. E. Dickerson (1994). *J. Mol. Biol.* **239**, 79–96.
	(B78) M. L. Kopka, D. S. Goodsell, K. Grzeskowiak, I. Baikalov, D. Cascio & R. E. Dickerson (1994). *Biochemistry*, **33**, 13593–13610.
	(B79) G. A. Leonard, K. E. McAuley-Hecht, N. J. Gibson, T. Brown, W. P. Watson & W. N. Hunter (1994). *Biochemistry*, **33**, 4755–4761.
	(B80) K. E. McAuley-Hecht, G. A. Leonard, N. J. Gibson, J. B. Thomson, W. P. Watson, W. N. Hunter & T. Brown (1994). *Biochemistry*, **33**, 10266–10270.
	(B81) C. M. Nunn, T. C. Jenkins & S. Neidle (1994). *Eur. J. Biochem.* **226**, 953–961.
	(B82) Z. Shakked, G. Guzlkevich-Guerstein, F. Frolow, D. Rabinovich, A. Joachimiak & P. B. Sigler (1994). *Nature (London)*, **368**, 469–473.
	(B83) N. Spink, D. G. Brown, J. V. Skelly & S. Neidle (1994). *Nucleic Acids Res.* **22**, 1607–1612.
	(B84) M. C. Vega, I. Garcia-Saez, J. Aymami, R. Eritja, G. A. van der Marel, J. H. van Boom, A. Rich & M. Coll (1994). *Eur. J. Biochem.* **222**, 721–726.
	(B85) Y. Timsit & D. Moras (1994). *EMBO J.* **13**, 2737–2746.
1995	(B86) K. Balendiran, S. T. Rao, C. Y. Sekharudu, G. Zon & M. Sundaralingam (1995). *Acta Cryst.* D**51**, 190–198.
	(B87) X. Chen, B. Ramakrishnan & M. Sundaralingam (1995). *Nature Struct. Biol.* **2**, 733–735.
	(B88) D. S. Goodsell, M. L. Kopka & R. E. Dickerson (1995). *Biochemistry*, **34**, 4983–4993.
	(B89) D. S. Goodsell, K. Grzeskowiak & R. E. Dickerson (1995). *Biochemistry*, **34**, 1022–1029.
	(B90) D. S. Goodsell, H. L. Ng, M. L. Kopka, J. W. Lown & R. E. Dickerson (1995). *Biochemistry*, **34**, 16654–16661.
	(B91) L. A. Lipscomb, M. E. Peek, M. L. Morningstar, S. M. Verghis, E. M. Miller, A. Rich, J. M. Essigmann & L. D. Williams (1995). *Proc. Natl Acad. Sci. USA*, **92**, 719–723.
	(B92) C. M. Nunn & S. Neidle (1995). *J. Med. Chem.* **38**, 2317–2325.
	(B93) L. W. Tari & A. S. Secco (1995). *Nucleic Acids Res.* **23**, 2065–2073.
	(B94) L. Van Meervelt, D. Vlieghe, A. Dautant, B. Gallois, G. Precigoux & O. Kennard (1995). *Nature (London)*, **374**, 742–744.
	(B95) J. Vojtechovsky, M. D. Eaton, B. Gaffney, R. Jones & H. M. Berman (1995). *Biochemistry*, **34**, 16632–16640.
	(B96) A. A. Wood, C. M. Nunn, A. Czarny, D. W. Boykin & S. Neidle (1995). *Nucleic Acids Res.* **23**, 3678–3684.
1996	(B97) T. J. Boggon, E. L. Hancox, K. E. McAuley-Hecht, B. A. Connolly, W. N. Hunter, T. Brown, R. T. Walker & G. A. Leonard (1996). *Nucleic Acids Res.* **24**, 951–961.
	(B98) G. R. Clark, E. J. Gray, S. Neidle, Y.-H. Li & W. Leupin (1996). *Biochemistry* **35**, 13745–13752.
	(B99) G. R. Clark, C. J. Squire, E. J. Gray, W. Leupin & S. Neidle (1996). *Nucleic Acids Res.* **24**, 4882–4889.
	(B100) C. A. Laughton, F. Tanious, C. M. Nunn, D. W. Boykin, W. D. Wilson & S. Neidle (1996). *Biochemistry*, **35**, 5655–5661.
	(B101) J. O. Trent, G. R. Clark, A. Kumar, W. D. Wilson, D. W. Boykin, J. E. Hall, R. R. Tidwell, B. L. Blagburn & S. Neidle (1996). *J. Med. Chem.* **39**, 4554–4562.
	(B102) L. Urpi, V. Tereshko, L. Malinina, T. Huynh-Dinh & J. A. Subirana (1996). *Nature Struct. Biol.* **3**, 325–328.
	(B103) D. Vlieghe, L. Van Meervelt, A. Dautant, B. Gallois, G. Precigoux & O. Kennard (1996). *Science*, **273**, 1702–1705.
	(B104) M. C. Wahl, S. T. Rao & M. Sundaralingam (1996). *Biophys. J.* **70**, 2857–2866.
1997	(B105) X. Chen, B. Ramakrishnan & M. Sundaralingam (1997). *J. Mol. Biol.* **267**, 1157–1170.
	(B106) G. R. Clark, D. W. Boykin, A. Czarny & S. Neidle (1997). *Nucleic Acids Res.* **25**, 1510–1515.
	(B107) G.-W. Han, M. L. Kopka, D. Cascio, K. Grzeskowiak & R. E. Dickerson (1997). *J. Mol. Biol.* **269**, 811–826.
	(B108) M. L. Kopka, D. S. Goodsell, G. W. Han, T. K. Chiu, J. W. Lown & R. E. Dickerson (1997). *Structure*, **5**, 1033–1046.
	(B109) C. Mayer-Jung, D. Moras & Y. Timsit (1997). *J. Mol. Biol.* **270**, 328–335.
	(B110) C. M. Nunn, E. Garman & S. Neidle (1997). *Biochemistry*, **36**, 4792–4799.

Table A23.4.1.2 (continued)

Year	Reference
	(B111) S. Portmann, K.-H. Altmann, N. Reynes & M. Egli (1997). *J. Am. Chem. Soc.* **119**, 2396–2403. (B112) H. Qiu, J. C. Dewan & N. C. Seeman (1997). *J. Mol. Biol.* **267**, 881–898. (B113) M. Shatzky-Schwartz, N. D. Arbuckle, M. Eisenstein, D. Rabinovich, A. Bareket-Samish, T. E. Haran, B. F. Luisi & Z. Shakked (1997). *J. Mol. Biol.* **267**, 565–623. (B114) A. A. Wood, C. M. Nunn, J. O. Trent & S. Neidle (1997). *J. Mol. Biol.* **269**, 827–841.

Table A23.4.1.3

X-ray analyses of Z helices

See introductory notes to Table A23.4.1.1. odm = 6*H*,8*H*-3,4-dihydropyrimido[4,5*c*][1,2]oxazin-7-one.

(*a*) Hexadecamers

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCGCGTTTTCGCGCG (hairpin)	$C2$	4	8	1988, UCLA	UDP011	(Z20, Z25)

(*b*) Decamers (disordered)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
GCGCGCGCGC	$P6_5$	6	2	1996, Ohio State	ZDJ050	(Z46)

(*c*) Octamers

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCICICG	$P6_5$	6	8	1992, Thomas Jefferson	ZDH030	(Z32)
CGCGCGCG	$P6_5$	6	8	1985, MIT	(ZDH017)	(Z10)
CGCATGCG	$P6_5$	6	8	1985, MIT	(ZDH016)	(Z10)

(*d*) Heptamers (overhanging 5′ bases)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
GCGCGCG	$P2_12_12_1$	4	6	1997, Oregon State	ZDG054	(Z50)
G^{5me}CGCGCG	$P2_12_12_1$	4	6	1997, Oregon State	ZDG055	(Z50)
GCGCGCG/ GCGCGCT	$P2_12_12_1$	4	6	1997, Oregon State	ZDG056	(Z50)
GCGCGCG	$P2_12_12_1$	4	6	1997, Ohio State	ZDG057	(Z51)

(*e*) Hexamers

(1) Alternating CG: Pu-Py alternation retained

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCGCG, Mg	$P2_12_12_1$	4	6	1989, MIT	ZDF002	(Z23)
CGCGCG, DL racemate	$P\bar{1}$	2	6	1993, Osaka	ZDF040	(Z36)
CGCGCG/spermine	$P2_12_12_1$	4	6	1991, MIT	ZDF029	(Z29)
CGCGCG/spermine, 163 K	$P2_12_12_1$	4	6	1994, MIT	ZDF035	(Z41)
CGCGCG/spermine, Mg	$P2_12_12_1$	4	6	1979, MIT	ZDF001	(Z1, Z23)
CGCGCG/spermidine	$P2_12_12_1$	4	6	1996, MIT	ZDF052	(Z47)
CGCGCG/thermospermidine	$P2_12_12_1$	4	6	1996, MIT	ZDF053	(Z48)
CGCGCG, Co, Mg	$P2_12_12_1$	4	6	1985, MIT	(ZDF019)	(Z11)
CGCGCG, Co, Mg	$P2_12_12_1$	4	6	1993, Illinois	(ZDF044)	(Z37)
CGCGCG/spermine, Co	$P2_12_12_1$	4	6	1993, Illinois	(ZDF045)	(Z37)
CGCGCG, Ru	$P2_12_12_1$	4	6	1987, MIT	(ZDF007)	(Z18)
CG c g CG	$P2_12_12_1$	4	6	1989, MIT	(ZHF026)	(Z24)

(2) Alternating CG: Pu-Py alternation broken

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CCGCGG	$C222_1$	8	6	1994, Moscow	UDF025	(Z42)

(3) Modified CG bases: Pu-Py alternation retained

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CG^{ara}CGCG	$P2_12_12_1$	4	6	1989, MIT	(ZDFS27)	(Z24)
CGC^{6mo}GCG	$P2_12_12_1$	4	6	1990, Rutgers	ZDFB21	(Z26)
CGCG^{4mo}CG	$P2_12_12_1$	4	6	1990, Cambridge	ZDFB25	(Z27)

Table A23.4.1.3 (continued)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCG^{4mo}CG	$P2_12_12_1$	4	6	1993, Manchester	ZDFB36	(Z35)
CGCG^{5br}CG	$P2_12_12_1$	4	6	1996, Manchester	ZDFB51	(Z49)
CGCG^{odm}CG	$P2_12_12_1$	4	6	1995, Cambridge	ZDFB43	(Z43)
^{5me}CG^{5me}CG^{5me}CG	$P2_12_12_1$	4	6	1982, MIT	ZDFB03	(Z6, Z7)
^{5br}CG^{5br}CG^{5br}CG, 291 K	$P2_12_12_1$	4	6	1986, Strasbourg	ZDFB04	(Z16, Z19)
^{5br}CG^{5br}CG^{5br}CG, 310 K	$P2_12_12_1$	4	6	1986, Strasbourg	ZDFB05	(Z16, Z19)
Aminohexyl-CG^{5br}CGCG	$C2$	4	6	1993, Illinois	(ZDFA32)	(Z38)
^{ara}CG^{ara}CG^{ara}CG (disordered)	$P6_522$	12	2	1992, Illinois	ZDFS33	(Z34)

(4) Modified CG bases: Pu-Py alternation broken

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
^{5me}CG$\underline{\mathrm{G}}$G^{5me}CG	$P2_12_12_1$	4	6	1993, Oregon State	ZDFB37	(Z40)

(5) With A, T, U, I bases: Pu-Py alternation retained

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
^{5me}CGTA^{5me}CG	$P2_12_12_1$	4	6	1984, MIT	ZDFB06	(Z8)
CGT^{2am}ACG	$P3_221$	6	3	1995, Rutgers	ZDFB41	(Z44)
CGT^{2am}ACG, Pt	$P3_221$	6	3	1995, Rutgers	ZDFB42	(Z44)
CGU^{2am}ACG	$P2_12_12_1$	4	6	1992, Rutgers	ZDFB31	(Z33)
^{5me}CGUA^{5me}CG	$P2_12_12_1$	4	6	1990, Oregon State	ZDFB24	(Z28)
^{5me}CGUA^{5me}CG, Cu	$P2_12_12_1$	4	5	1991, Oregon State	ZDFB10	(Z30)
CACGTG	$P2_12_12_1$	4	6	1988, MIT	(ZDF008)	(Z21)
C^{2am}ACGTG	$P2_12_12_1$	4	6	1986, MIT	ZDFB11	(Z17)
CGCICG	$P2_12_12_1$	4	6	1993, Thomas Jefferson	(ZDFB34)	(Z39)
CACGCG/CGCGTG	$P2_12_12_1$	4	6	1995, Madras	ZDF039	(Z45)
CGCACG/CGTGCG	$P2_1$	2	6	1995, Madras	ZDF038	(Z45)

(6) With A, T, U, I bases: Pu-Py alternation broken

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
^{5br}CG$\underline{\mathrm{AT}}$^{5br}CG	$P2_12_12_1$	4	6	1985, MIT	(ZDFB09)	(Z13)

(7) With mismatches (underlined)

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
C$\underline{\mathrm{G}}$CG$\underline{\mathrm{T}}$G	$P2_12_12_1$	4	6	1985, MIT	ZDF013	(Z12)
C$\underline{\mathrm{G}}$CG^{5fl}$\underline{\mathrm{U}}$G	$P2_12_12_1$	4	6	1989, MIT	ZDFB12	(Z22)
^{5br}$\underline{\mathrm{U}}$GCGC$\underline{\mathrm{G}}$	$P2_12_12_1$	4	6	1986, Cambridge	ZDFB14	(Z15)
C$\underline{\mathrm{G}}$CG$\underline{\mathrm{T}}$G, Co, Mg	$P2_12_12_1$	4	6	1993, Illinois	ZDF046	(Z37)
C$\underline{\mathrm{G}}$CG$\underline{\mathrm{T}}$G, Cu, Mg	$P2_12_12_1$	4	6	1993, Illinois	ZDF047	(Z37)
^{5me}C$\underline{\mathrm{G}}$^{5me}CG$\underline{\mathrm{T}}$G, Ba	$P2_12_12_1$	4	6	1993, Illinois	(ZDFB48)	(Z37)

(*f*) Tetramers

Sequence	Space group	*Z*	Ubp	Date, institution	NDB No.	Reference
CGCG	$C222_1$	8	4	1980, UCLA (CIT)	ZDD015	(Z3, Z4, Z5)
CGCG (disordered)	$P6_5$	6	6	1980, MIT	ZDD023	(Z2)

References (numbered chronologically by year and alphabetically by first author within each year)

Year	Reference
1979	(Z1) A. H.-J. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel & A. Rich (1979). *Nature (London)*, **282**, 680–686.
1980	(Z2) J. L. Crawford, F. J. Kolpak, A. H.-J. Wang, G. J. Quigley, J. H. van Boom, G. van der Marel & A. Rich (1980). *Proc. Natl Acad. Sci. USA*, **77**, 4016–4020.
	(Z3) H. R. Drew, T. Takano, S. Tanaka, K. Itakura & R. E. Dickerson (1980). *Nature (London)*, **286**, 567–573.
1981	(Z4) R. E. Dickerson, H. R. Drew & B. N. Conner (1981). *Biomolecular stereodynamics*, Vol. 1, edited by R. H. Sarma, pp. 1–34. New York: Adenine Press.
	(Z5) H. R. Drew & R. E. Dickerson (1981). *J. Mol. Biol.* **152**, 723–736.
1982	(Z6) S. Fujii, A. H.-J. Wang, G. van der Marel, J. H. van Boom & A. Rich (1982). *Nucleic Acids Res.* **10**, 7879–7892.
	(Z7) S. Fujii, A. H.-J. Wang, J. van Boom & A. Rich (1982). *Nucleic Acids Res. Symp. Ser.* **11**, 109–112.
1984	(Z8) A. H.-J. Wang, T. Hakoshima, G. van der Marel, J. H. van Boom & A. Rich (1984). *Cell*, **37**, 321–331.
1985	(Z9) R. G. Brennan & M. Sundaralingam (1985). *J. Mol. Biol.* **181**, 561–563.
	(Z10) S. Fujii, A. H.-J. Wang, G. J. Quigley, H. Westerink, G. van der Marel, J. H. van Boom & A. Rich (1985). *Biopolymers*, **24**, 243–250.
	(Z11) R. V. Gessner, G. J. Quigley, A. H.-J. Wang, G. A. van der Marel, J. H. van Boom & A. Rich (1985). *Biochemistry*, **24**, 237–240.
	(Z12) P. S. Ho, C. A. Frederick, G. J. Quigley, G. A. van der Marel, J. H. van Boom, A. H.-J. Wang & A. Rich (1985). *EMBO J.* **4**, 3617–3623.
	(Z13) A. H.-J. Wang, R. V. Gessner, G. A. van der Marel, J. H. van Boom & A. Rich (1985). *Proc. Natl Acad. Sci. USA*, **82**, 3611–3615.

Table A23.4.1.3 (continued)

Year	Reference
1986	(Z14) R. G. Brennan, E. Westhof & M. Sundaralingam (1986). *J. Biomol. Struct. Dyn.* **3**, 649–665.
	(Z15) T. Brown, G. Kneale, W. N. Hunter & O. Kennard (1986). *Nucleic Acids Res.* **14**, 1801–1809.
	(Z16) B. Chevrier, A. C. Dock, B. Hartmann, M. Leng, D. Moras, M. T. Thuong & E. Westhof (1986). *J. Mol. Biol.* **188**, 707–719.
	(Z17) M. Coll, A. H.-J. Wang, G. A. van der Marel, J. H. van Boom & A. Rich (1986). *J. Biomol. Struct. Dyn.* **4**, 157–172.
1987	(Z18) P. S. Ho, C. A. Frederick, D. Saal, A. H.-J. Wang & A. Rich (1987). *J. Biomol. Struct. Dyn.* **4**, 521–534.
	(Z19) E. Westhof (1987). *J. Biomol. Struct. Dyn.* **5**, 581–600.
1988	(Z20) R. Chattopadhyaya, S. Ikuta, K. Grzeskowiak & R. E. Dickerson (1988). *Nature (London)*, **334**, 175–179.
	(Z21) M. Coll, I. Fita, J. Lloveras, J. A. Subirana, F. Bardella, T. Huynh-Dinh & J. Igolen (1988). *Nucleic Acids Res.* **16**, 8695–8705.
1989	(Z22) M. Coll, D. Saal, C. A. Frederick, J. Aymami, A. Rich & A. H.-J. Wang (1989). *Nucleic Acids Res.* **17**, 911–923.
	(Z23) R. V. Gessner, C. A. Frederick, G. J. Quigley, A. Rich & A. H.-J. Wang (1989). *J. Biol. Chem.* **264**, 7921–7935.
	(Z24) M.-K. Teng, Y.-C. Liaw, G. A. van der Marel, J. H. van Boom & A. H.-J. Wang (1989). *Biochemistry*, **28**, 4923–4928.
1990	(Z25) R. Chattopadhyaya, K. Grzeskowiak & R. E. Dickerson (1990). *J. Mol. Biol.* **211**, 189–210.
	(Z26) S. L. Ginell, S. Kuzmich, R. A. Jones & H. M. Berman (1990). *Biochemistry*, **29**, 10461–10465.
	(Z27) L. Van Meervelt, M. H. Moore, P. K. T. Lin, D. M. Brown & O. Kennard (1990). *J. Mol. Biol.* **216**, 773–781.
	(Z28) G. Zhou & P. S. Ho (1990). *Biochemistry*, **29**, 7229–7236.
1991	(Z29) M. Egli, L. D. Williams, Q. Gao & A. Rich (1991). *Biochemistry*, **30**, 11388–11402.
	(Z30) B. H. Geierstanger, T. F. Kagawa, S.-L. Chen, G. J. Quigley & P. S. Ho (1991). *J. Biol. Chem.* **266**, 20185–20191.
1992	(Z32) V. D. Kumar, R. W. Harrison, L. C. Andrews & I. T. Weber (1992). *Biochemistry*, **31**, 1541–1550.
	(Z33) B. Schneider, S. L. Ginnell, R. Jones, B. Gaffney & H. M. Berman (1992). *Biochemistry*, **21**, 9622–9628.
	(Z34) H. Zhang, G. A. van der Marel, J. H. van Boom & A. H.-J. Wang (1992). *Biopolymers*, **32**, 1559–1569.
1993	(Z35) A. R. Cervi, A. Guy, G. A. Leonard, R. Teoule & W. N. Hunter (1993). *Nucleic Acids Res.* **21**, 5623–5629.
	(Z36) M. Doi, M. Inoue, K. Tomoo, T. Ishida, Y. Ueda, M. Akagi & H. Urata (1993). *J. Am. Chem. Soc.* **115**, 10432–10433.
	(Z37) Y.-G. Gao, K. Sriram & A. H.-J. Wang (1993). *Nucleic Acids Res.* **21**, 4093–4101.
	(Z38) Y.-C. Jean, Y.-G. Gao & A. H.-J. Wang (1993). *Biochemistry*, **32**, 381–388.
	(Z39) V. D. Kumar & I. T. Weber (1993). *Nucleic Acids Res.* **9**, 2201–2208.
	(Z40) G. P. Schroth, T. F. Kagawa & P. Shing Ho (1993). *Biochemistry*, **32**, 13381–13392.
1994	(Z41) D. Bancroft, L. D. Williams, A. Rich & M. Egli (1994). *Biochemistry*, **33**, 1073–1086.
	(Z42) L. Malinina, L. Urpi, X. Salas, T. Huynh-Dinh & J. A. Subirana (1994). *J. Mol. Biol.* **243**, 484–493.
1995	(Z43) M. H. Moore, L. Van Meervelt, S. A. Salisbury, P. Kong Thoo Lin & D. M. Brown (1995). *J. Mol. Biol.* **251**, 665–673.
	(Z44) G. N. Parkinson, G. M. Arvantis, L. Lessinger, S. L. Ginnell, R. Jones, B. Gaffney & H. M. Berman (1995). *Biochemistry*, **34**, 15487–15495.
	(Z45) C. Sadasivan & N. Gautham (1995). *J. Mol. Biol.* **248**, 918–930.
1996	(Z46) C. Ban, B. Ramakrishnan & M. Sundaralingam (1996). *Biophys. J.* **71**, 1215–1221.
	(Z47) H. Ohishi, I. Nakanishi, K. Inubushi, G. A. van der Marel, J. H. van Boom, A. Rich, A. H.-J. Wang, T. Hakoshima & K. Tomita (1996). *FEBS Lett.* **391**, 143–156.
	(Z48) H. Ohishi, N. Terasoma, I. Nakanishi, G. A. van der Marel, J. H. van Boom, A. Rich, A. H.-J. Wang, T. Hakoshima & K. Tomita (1996). *FEBS Lett.* **398**, 291–296.
	(Z49) M. R. Peterson, S. J. Harrop, S. M. McSweeney, G. A. Leonard, A. W. Thompson, W. N. Hunter & J. R. Helliwell (1996). *J. Synchrotron Rad.* **3**, 24–34.
1997	(Z50) B. H. M. Mooers, B. F. Eichman & P. S. Ho (1997). *J. Mol. Biol.* **269**, 796–810.
	(Z51) B. Pan, C. Ban, M. Wahl & M. Sundaralingam (1997). *Biophys. J.* **83**, 1553–1561.

References

Altona, C., Geise, H. J. & Romers, C. (1968). *Conformation of non-aromatic ring compounds, XXIV. On the geometry of the perhydrophenanthrene skeleton in some steroids. Tetrahedron*, **24**, 13–32.

Altona, C. & Sundaralingam, M. (1972). *Conformational analysis of the sugar ring in nucleosides and nucleotides. J. Am. Chem. Soc.* **94**, 8205–8212.

Ansevin, A. T. & Wang, A. H. (1990). *Evidence for a new Z-type left-handed DNA helix. Nucleic Acids Res.* **18**, 6119–6126.

Arnott, S. (1970). *The geometry of nucleic acids. Prog. Biophys. Mol. Biol.* **21**, 265–319.

Babcock, M. S. & Olson, W. K. (1994). *The effect of mathematics and coordinate system on comparability and 'dependencies' of nucleic acid structure parameters. J. Mol. Biol.* **237**, 98–124.

Babcock, M. S., Pednault, E. & Olson, W. (1993). *Nucleic acid structure analysis: a users guide to a collection of new analysis programs. J. Biomol. Struct. Dyn.* **11**, 597–628.

Babcock, M. S., Pednault, E. & Olson, W. (1994). *Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. J. Mol. Biol.* **237**, 125–156.

Basham, B., Eichman, B. F. & Ho, P. S. (1998). *The single-crystal structures of Z-DNA*. In *Oxford Handbook of Nucleic Acid Structure*, edited by S. Neidle, ch. 7, pp. 200–252. Oxford University Press.

Berman, H. M. (1996). *Crystal studies of B-DNA: the answers and the questions. Biopolymers Nucleic Acid Sci.* **44**, 23–44.

Bugg, C. E., Thomas, J. M., Sundaralingam, M. & Rao, S. T. (1971). *Stereochemistry of nucleic acids and their constituents. X. Solid-state base-stacking patterns in nucleic acid consituents and polynucleotides. Biopolymers*, **10**, 175–219.

Crick, F. H. C. & Watson, J. D. (1954). *The complementary structure of deoxyribonucleic acid. Proc. R. Soc. London Ser. A*, **223**, 80–96.

Crothers, D. M. & Drak, J. (1992). *Global features of DNA structure by comparative gel electrophoresis. Methods Enzymol.* **212**, 46–71.

Crothers, D. M., Haran, T. E. & Nadeau, J. G. (1990). *Intrinsically bent DNA. J. Biol. Chem.* **265**, 7093–7096.

Davies, D. B. (1978). *Conformations of nucleosides and nucleotides. Prog. Nucl. Magn. Reson. Spectros.* **12**, 135–186.

Dickerson, R. E. (1972). *The structure and history of an ancient protein. Sci. Am.* **226** (April), 58–72.

Dickerson, R. E. (1983). *The DNA helix and how it is read. Sci. Am.* **249** (December), 94–111.

Dickerson, R. E. (1992). *DNA structure from A to Z. Methods Enzymol.* **211**, 67–111.

Dickerson, R. E. (1997*a*). *Obituary: Irving Geis, 1908–1997. Structure*, **5**, 1247–1249.

Dickerson, R. E. (1997*b*). *Irving Geis, molecular artist, 1908–1997. Protein Sci.* **6**, 2843–2844.

Dickerson, R. E. (1997*c*). *Biology in pictures: molecular artistry. Curr. Biol.* **7**, R720–R741.

Dickerson, R. E. (1998*a*). *Sequence-dependent B-DNA conformation in crystals and in protein complexes*. In *Structure, Motion, Interaction and*

Expression of Biological Macromolecules, edited by R. H. Sarma & M. H. Sarma, pp. 17–36. New York: Adenine Press.

Dickerson, R. E. (1998*b*). *Helix structure and molecular recognition by B-DNA*. In *Oxford Handbook of Nucleic Acid Structure*, edited by S. Neidle, ch. 7, pp. 145–197. Oxford University Press.

Dickerson, R. E. (1998*c*). *DNA bending: the prevalence of kinkiness and the virtues of normality. Nucleic Acids Res.* **26**, 1906–1926.

Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., Lavery, R., Nelson, H. C. M., Olson, W. K., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D. M., Tung, C.-S., von Kitzing, E., Wang, A. H.-J. & Zhurkin, V. B. (1989). *Definitions and nomenclature of nucleic acid structure components. EMBO J.* **8**, 1–4; *J. Biomol. Struct. Dyn.* **6**, 627–634; *Nucleic Acids Res.* **17**, 1797–1803; *J. Mol. Biol.* **206**, 787–791.

Dickerson, R. E. & Chiu, T. K. (1997). *Helix bending as a factor in protein/DNA recognition. Biopolymers Nucleic Acid Sci.* **44**, 361–403.

Dickerson, R. E. & Geis, I. (1969). *The Structure and Action of Proteins*. New York: Harper & Row and Menlo Park: W. A. Benjamin Co.

Dickerson, R. E. & Geis, I. (1976). *Chemistry, Matter and the Universe*. Menlo Park: Benjamin/Cummings Co.

Dickerson, R. E. & Geis, I. (1983). *Hemoglobin: Structure, Function, Evolution, and Pathology*. Menlo Park: Benjamin/Cummings Co.

Dickerson, R. E., Goodsell, D. & Kopka, M. L. (1996). *MPD and DNA bending in crystals and in solution. J. Mol. Biol.* **256**, 108–125.

Dickerson, R. E., Goodsell, D. S., Kopka, M. L. & Pjura, P. E. (1987). *The effect of crystal packing on oligonucleotide double helix structure. J. Biomol. Struct. Dyn.* **5**, 557–579.

Dickerson, R. E., Goodsell, D. S. & Neidle, S. (1994). *… the tyranny of the lattice …. Proc. Natl Acad. Sci. USA*, **91**, 3579–3583.

El Hassan, M. A. & Calladine, C. R. (1997). *Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. Philos. Trans. R. Soc. London A*, **355**, 43–100.

Feigon, J. (1996). *DNA triplexes, quadruplexes & aptamers*. In *Encyclopedia of Nuclear Magnetic Resonance*, edited by D. M. Grant & R. K. Harris, pp. 1726–1731. New York: Wiley.

Franklin, R. E. & Gosling, R. G. (1953). *The structure of sodium thymonucleate fibres. I. The influence of water content. Acta Cryst.* **6**, 673–677.

Haschmeyer, A. E. V. & Rich, A. (1967). *Nucleoside conformation: an analysis of steric barriers to rotation about the glycosidic bond. J. Mol. Biol.* **27**, 369–384.

Herbert, A. & Rich, A. (1996). *The biology of left-handed Z-DNA. J. Biol. Chem.* **271**, 11595–11598.

Ho, P. S. & Mooers, B. H. M. (1996). *Z-DNA crystallography. Biopolymers Nucleic Acid Sci.* **44**, 65–90.

Hoogsteen, K. (1963). *The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. Acta Cryst.* **16**, 907–916.

Hunter, C. A. & Sanders, J. K. M. (1990). *The nature of π–π interactions. J. Am. Chem. Soc.* **112**, 5525–5534.

Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J. & Dickerson, R. E. (1996). *How proteins recognize the TATA box. J. Mol. Biol.* **261**, 239–254.

Kendrew, J. C. (1961). *The three-dimensional structure of a protein molecule. Sci. Am.* **205** (December), 96–110.

Kim, J. L., Nikolov, D. B. & Burley, S. K. (1993). *Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature (London)*, **365**, 520–527.

Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. (1993). *Crystal structure of a yeast TBP/TATA-box complex. Nature (London)*, **365**, 512–520.

Koo, H.-S., Drak, J., Rice, J. A. & Crothers, D. M. (1990). *Determination of the extent of DNA bending by an adenine-thymine tract. Biochemistry*, **29**, 4227–4234.

Koo, H.-S., Wu, H.-M. & Crothers, D. M. (1986). *DNA bending at adenine-thymine tracts. Nature (London)*, **320**, 501–506.

Kostrewa, D. & Winkler, F. K. (1995). *Mg^{2+} binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. Biochemistry*, **34**, 683–696.

Langridge, R., Marvin, D. A., Seeds, W. E., Wilson, H. R., Hooper, C. W., Wilkins, M. H. F. & Hamilton, L. D. (1960). *The molecular configurations of deoxyribonucleic acid. II. Molecular models and their Fourier transforms. J. Mol. Biol.* **2**, 38–64.

Lavery, R. & Sklenar, H. (1988). *The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. J. Biomol. Struct. Dyn.* **6**, 63–91.

Lavery, R. & Sklenar, H. (1989). *Defining the structure of irregular nucleic acids: conventions and principles. J. Biomol. Struct. Dyn.* **6**, 655–667.

Leslie, A. G. W., Arnott, S., Chandrasekaran, R. & Ratliff, R. L. (1980). *Polymorphism of DNA double helices. J. Mol. Biol.* **143**, 49–72.

Levitt, M. & Warshel, A. (1978). *Extreme conformational flexibility of the furanose ring in DNA and RNA. J. Am. Chem. Soc.* **100**, 2607–2613.

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G. & Lu, P. (1996). *Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science*, **271**, 1247–1254.

Marini, J. C., Levene, S. D., Crothers, D. M. & Englund, P. T. (1982). *Bent helical structures in kinetoplast DNA. Proc. Natl Acad. Sci. USA*, **79**, 7664–7668.

Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G. & Burley, S. K. (1996). *Crystal structure of a human TATA box-binding protein/TATA element complex. Proc. Natl Acad. Sci. USA*, **93**, 4862–4867.

Parkinson, G., Wilson, C., Gunasekera, A., Ebright, Y. W., Ebright, R. H. & Berman, H. M. (1996). *Structure of the CAP–DNA complex at 2.5 angstroms resolution: a complete picture of the protein–DNA interface. J. Mol. Biol.* **260**, 395–408.

Pelton, J. G. & Wemmer, D. E. (1989). *Structural characterization of a 2:1 distamycin A/d(CGCAAATTGGC) complex by two-dimensional NMR. Proc. Natl Acad. Sci. USA*, **86**, 5723–5727.

Pelton, J. G. & Wemmer, D. E. (1990). *Binding modes of distamycin-A with d(CGCAAATTTGCG)2 determined by two-dimensional NMR. J. Am. Chem. Soc.* **112**, 1393–1399.

Phillips, D. C. (1966). *The three-dimensional structure of an enzyme molecule. Sci. Am.* **215** (November), 78–90.

Pohl, F. M. (1976). *Polymorphism of a synthetic DNA in solution. Nature (London)*, **260**, 365–366.

Pohl, F. M. & Jovin, T. M. (1972). *Salt-induced co-operative conformational change of a syhnthetic DNA: equlibrium and kinetic studies with poly(dG-dC). J. Mol. Biol.* **67**, 375–396.

Rice, P. A., Yang, S.-W., Mizuuchi, K. & Nash, H. A. (1996). *Crystal structure of an IHF–DNA complex: a protein-induced DNA U-turn. Cell*, **87**, 1295–1306.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*. New York, Berlin, Heidelberg and Tokyo: Springer-Verlag.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Conformations of the sugar–phosphate backbone in helical DNA crystal structures. Biopolymers*, **42**, 113–124.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). *Crystal structure of a CAP–DNA complex: the DNA is bent by 90 degrees. Science*, **253**, 1001–1007.

Schumacher, M. A., Choi, K. Y., Zalkin, H. & Brennan, R. G. (1994). *Crystal structure of LacI member, PurR, bound to DNA: minor groove binding by alpha helices. Science*, **266**, 763–770.

Schwartz, T., Rould, M. A., Lowenjaupt, K., Herbert, A. & Rich, A. (1999). *Crystal structure of the Zα domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. Science*, **284**, 1841–1845.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). *Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl Acad. Sci. USA*, **73**, 804–808.

Sklenár, V. & Feigon, J. (1990). *Formation of a stable triplex from a single DNA strand. Nature (London)*, **345**, 836–838.

Sprous, D., Young, M. A. & Beveridge, D. L. (1999). *Molecular dynamics studies of axis bending in d(G_5-(GA_4T_4C)$_2$-C_5) and d(G_5-(GT_4A_4C)$_2$-C5): effects of sequence polarity on DNA curvature. J. Mol. Biol.* **285**, 1623–1632.

Sprous, D., Zacharias, W., Wood, Z. A. & Harvey, S. C. (1995). *Dehydrating agents sharply reduce curvature in DNAs containing A-tracts. Nucleic Acids Res.* **23**, 1816–1821.

Sundaralingam, M. (1975). *Principles governing nucleic acid and polynucleotide conformations*. In *Structure and Conformation of Nucleic Acids and Protein–Nucleic Acid Interactions*, edited by M. Sundaralingam & S. T. Rao, pp. 487–524. Baltimore: University Park Press.

Thomas, K. A., Smith, G. M., Thomas, T. B. & Feldmann, R. J. (1982). *Electronic distributions within protein phenylalanine aromatic rings are reflected by the three-dimensional oxygen atom environments. Proc. Natl Acad. Sci. USA*, **79**, 4843–4847.

Voet, D. & Voet, J. G. (1990). *Biochemistry*. New York: John Wiley & Sons.

Voet, D. & Voet, J. G. (1995). *Biochemistry*, 2nd edition. New York: John Wiley & Sons.

Wahl, M. C. & Sundaralingam, M. (1996). *Crystal structures of A-DNA duplexes. Biopolymers Nucleic Acid Sci.* **44**, 45–63.

Wahl, M. C. & Sundaralingam, M. (1998). *A-DNA duplexes in the crystal*. In *Oxford Handbook of Nucleic Acid Structure*, edited by S. Neidle, ch. 5, pp. 117–144. Oxford University Press.

Watson, J. D. & Crick, F. H. C. (1953). *Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature (London)*, **171**, 737–738.

Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P., Bryan, R. K., Martin, P. D., Petratos, K. & Wilson, K. S. (1993). *The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. EMBO J.* **12**, 1781–1795.

Wu, H.-M. & Crothers, D. M. (1984). *The locus of sequence-directed and protein-induced DNA bending. Nature (London)*, **308**, 509–513.

Yang, W. & Steitz, T. A. (1995). *Crystal structure of the site-specific recombinase gamma delta resolvase complexed with a 34 bp cleavage site. Cell*, **82**, 193–207.

International Tables for Crystallography (2012). Vol. F, Chapter 23.5, pp. 800–820.

Chapter 23.5. Solvent structure

C. Mattos and D. Ringe

23.5.1. Introduction

The unique properties of water and its role in nature have preoccupied the minds of scientists and philosophers for centuries. However, only relatively recently have the tools become available to study the specific roles that water molecules play with respect to protein structure and function. When the first crystal structure of a protein was obtained by X-ray diffraction (Kendrew, 1963), the focus was on the arrangement of the amino-acid residues into secondary and tertiary structure. Although the presence of water molecules associated with the protein was noticed, little attention was given to their structure and possible functional role. The structure of the protein itself was a great novelty, and its features were eagerly analysed. For many years, the crucial role of water molecules in maintaining both the structural integrity and the functional viability of proteins was not completely obvious, although in the 1950s Kauzmann argued correctly that water plays an important role in maintaining protein structure (Kauzmann, 1959). In the late 1970s and early 1980s, reviews began to appear focusing on the properties of water relevant to interaction with proteins (Edsall & McKenzie, 1978) and the location and role of water molecules on protein surfaces (Blake *et al.*, 1983; Edsall & McKenzie, 1983). As high-resolution structures became more easily attainable and refinement techniques improved, the importance of water molecules became increasingly apparent, and solvent structure now occupies a front seat in the realm of structural biology. There is a strong sense in the scientific community that water molecules play an integral role in many aspects of protein structure and function, and great effort is now being focused on understanding solvent effects in precise atomic detail.

In principle, water molecules can contribute both enthalpically and entropically to any process in which they are involved. The main contribution of water as solvent in the protein-folding process, for example, is entropic, driving the collapse of hydrophobic residues into the core of the protein. As is currently understood, the general shape of globular proteins is attained by this effect, with the specific structural features guided by the hydrogen bonds that define secondary structure (Hendsch & Tidor, 1994; Hendsch *et al.*, 1996). Although the solvent contribution to the protein-folding process is beyond the scope of this chapter, its essential role has been recently reviewed elsewhere (Mattos & Clark, 2008). It is important to understand protein three-dimensional structures as having evolved in bulk water, a fact which is largely invisible to the current methods in structural biology. The size, geometry, planarity and orientational flexibility of water molecules give them structural and functional importance. All folded globular proteins have evolved with tens to hundreds of binding sites specific for water molecules, a situation contrary to that of larger ligands which usually bind in a single or small number of specific sites found in a given protein or family of proteins. In this respect, water is unique and its ubiquitous appearance is not a direct consequence of its chemical properties alone, but has an evolutionary origin. Proteins evolved in an aqueous milieu, where over time some water molecules were specifically incorporated as integral parts of the protein architecture.

At first glance, the surface of a protein determined by X-ray crystallography appears randomly populated by a layer of water molecules. A careful analysis, however, reveals that the arrangement of water molecules on protein surfaces is not random. In folded proteins, individual water molecules participate in a variety of structural and functional roles, ranging from filling small cavities that are not fully occupied by protein atoms to allowing flexibility, such as in the case of charged surface side chains that can move freely while continuously maintaining hydrogen-bonding partners. Water molecules can fill deep crevices on the protein surface, or they can play a crucial role in the thermodynamics of ligand binding. The mobility as well as the number and strength of hydrogen-bonding partners that are observed for water molecules bound to protein surfaces vary considerably, and it is becoming increasingly apparent that these factors are correlated with functional roles. The atomic coordinates for any protein should not be considered complete without those bound solvent molecules that can be observed, for they are part of the structure.

Bound water molecules have been implicated and studied in the context of substrate specificity and affinity (Quiocho *et al.*, 1989; Herron *et al.*, 1994; Ladbury, 1996), catalysis (Privé *et al.*, 1992; Singer *et al.*, 1993; Komives *et al.*, 1995), mediation of protein–DNA interactions (Clore *et al.*, 1994; Shakked *et al.*, 1994; Morton & Ladbury, 1996), cooperativity (Royer *et al.*, 1996), conformational stability (Bhat *et al.*, 1994), and drug design (Poormina & Dean, 1995*a*,*b*,*c*). One of the challenges now is to translate the structural information observed into a thermodynamic understanding of the water contribution to the various processes. In some cases, an attempt has been made to relate changes in water structure between two forms of a protein (*e.g.* ligated and unligated or native and mutant) to changes in the measured heat capacity (Holdgate *et al.*, 1997) or to measurements of enthalpy and entropy changes by titration calorimetry (Bhat *et al.*, 1994). Thermodynamic solvent isotope effects have also been reported, where the thermodynamics of association of several binding processes were evaluated calorimetrically in light and heavy water (Chervenak & Toone, 1994). In other cases, the three-dimensional structures were directly interpreted in terms of thermodynamic contributions (Quiocho *et al.*, 1989; Morton & Ladbury, 1996). Ultimately, a thorough understanding of the thermodynamics and kinetics underlying solvent structure will lead to powerful predictive methods. Theoreticians, on the one hand, have developed models based on physical principles and use experimental knowledge to assess whether their predictions are correct. Experimentalists, on the other hand, attempt to explain the observed phenomena in terms of the well established physical theories that govern the natural world. Progress is being made on both fronts, but a large gap still remains between the two. A bridge is being built from both sides of the gap and when the two sides meet at a common point, the many pieces of this complicated puzzle will have been deciphered and put in their proper places, so that a global view of molecular processes in

Copyright © 2012 International Union of Crystallography

water can be obtained from whatever perspective one wishes to take: chemical, physical or biological.

The present chapter summarizes the empirical information gathered over the years on the structure of water molecules bound to proteins. The focus will be on structures solved by X-ray crystallography, although complementary techniques of obtaining solvent structure will be discussed briefly and, when appropriate, particular examples will be given. Section 23.5.2 is concerned with the methods by which solvent structure can be observed, Section 23.5.3 summarizes knowledge derived from database analysis of large numbers of proteins, Section 23.5.4 focuses on particular examples of groups of well studied protein structures, Section 23.5.5 discusses the contribution of protein models obtained at very high resolution to the understanding of solvent structure, and Section 23.5.6 contains an analysis of water molecules as mediators of complex formation. Finally, Section 23.5.7 presents a conclusion and a perspective regarding the direction in which this information can lead in building a cohesive understanding of the roles played by solvent in the structural integrity and biological function of macromolecules.

23.5.2. Determination of water molecules

The most prominent method by which the structure of water molecules on the surface of macromolecules can be observed at the atomic level is X-ray crystallography. The information classically available from this methodology is on bound water molecules, characterized by a high probability density and reduced mobility relative to the bulk solvent, which results in clearly observed electron density. Information on water structure at larger distances from the protein is available in the low-resolution reflections, but this information is more difficult to capture. This chapter focuses on the water molecules for which there is information at high resolution (>3 Å), although great progress has been made in recent years in modelling the disordered water structure at the protein–solvent interface, enabling more effective use of the low-resolution data (Badger, 1993; Jiang & Brünger, 1994; Lounnas *et al.*, 1994). Typically, one is interested in studying solvent structure because of the effects that it has on the protein. Lounnas *et al.* (1994) gave a particularly interesting focus on the effect of the protein on the solvent structure surrounding it. Using a combination of molecular-dynamics simulations of explicitly solvated myoglobin and the low resolution X-ray data from myoglobin crystals, they devised a method to describe the effect of the protein on the solvent structure to a distance of 6 Å from the surface. They found that the mobility and probability density of water molecules perpendicular to the protein surface varied considerably depending on the particular composition and three-dimensional structure of the amino-acid residues at the particular area of interest (Lounnas & Pettitt, 1994; Lounnas *et al.*, 1994).

There are a variety of criteria that have been used in placing crystallographic water molecules in electron-density maps. For tightly bound water molecules, with low *B* factors, the placement involves little or no subjectivity, but the choice of whether or not to include the more disordered waters (or those with low occupancy) can be rather subjective. It generally involves picking the electron-density contour level and *B*-factor cutoffs as well as making a choice of whether to use a simple difference electron-density map ($F_o - F_c$) or to use a higher-order difference electron-density map ($2F_o - F_c$ or $3F_o - 2F_c$). One criterion, applied consistently in placing water molecules on the surface of elastase structures, is the simultaneous presence of electron density at the 3σ contour level in an $F_o - F_c$ electron-density map and at the 1σ contour level in a $2F_o - F_c$ electron-density map. After refinement, those waters are kept that have a *B* factor of 50 Å^2 or less. A few exceptions do occur, where there is clear electron density for a water molecule with a *B* factor of up to 60 Å^2. Virtually all of the water molecules placed by these criteria have at least one hydrogen bond to a protein atom or to another water molecule and are mainly part of the first hydration shell on the protein surface.

A method that has provided information on solvent structure complementary to that obtained by X-ray crystallography is based on $D_2O - H_2O$ neutron difference maps (Shpungin & Kossiakoff, 1986). The main advantage of this methodology is in locating partially ordered water molecules whose electron-density peaks may be at the limit of the signal-to-noise ratio allowed for confidently determining positions of water molecules by X-ray diffraction. Scattering of neutrons by H_2O and D_2O is quite different, while scattering from the protein remains the same. Therefore, difference maps based on the two data sets should average to zero where the protein is present and result in peaks only where water molecules are found. Neutron scattering is particularly suited to this because of the threefold greater scattering power of deuterated water molecules relative to light water, providing a larger signal-to-noise ratio in assigning water positions. This method is particularly useful in detecting the second hydration sphere on protein surfaces (Kossiakoff *et al.*, 1992).

NMR spectroscopy can also serve as a complementary technique, providing dynamic information on the lifetime of interaction of a single water molecule on the protein surface. The fact that, with few exceptions, no cross-relaxation peaks are observed at the protein–water interface is an indication that the motion timescale for water molecules in contact with protein is close to that in bulk water at room temperature. The NMR data suggest that water molecules observed in crystal structures have lifetimes of the order of tens of nanoseconds or less (Bryant, 1996). A small number of relatively long-lived structural waters (with residence times in the range 10^{-2} to 10^{-8} s) can be detected by modern NMR techniques (Otting *et al.*, 1991). Four water molecules have been detected by NMR in bovine pancreatic trypsin inhibitor (BPTI) (Otting & Wuthrich, 1989) and six have been observed in complexes of human dihydrofolate reductase with methotrexate (Meiering & Wagner, 1995). Observation of these waters in the corresponding crystal structures reveals that they are tightly bound waters, with three or four hydrogen bonds to protein atoms, and many are found to bridge between secondary-structure elements or are found to mediate protein–ligand interaction (Meiering & Wagner, 1995). It is important to understand, then, that water molecules near protein surfaces occupy energy minima favoured by hydrogen bonding and ion–dipole effects, which results in water molecules being present in these positions more often than in others. Although when looking at a crystallographic protein structure it is easy to think of a given site as being occupied by a single water molecule, it is in fact only the site that is single, with an enormous number of different individual water molecules sampling it during the time of data collection. This was qualitatively understood from the beginning, but NMR experiments have played a key role in setting quantitative upper boundaries to the residence times of water molecules on the protein surface.

Finally, mention must be made of the computational efforts invested in representing and understanding solvent structure on macromolecular surfaces. The computational work encompasses a variety of methodologies, including integral equation methods (Beglov & Roux, 1997), molecular dynamics (Brooks & Karplus, 1989; Hayward *et al.*, 1993; van Gunsteren *et al.*, 1994), thermodynamic understanding through free-energy simulations (Roux *et al.*, 1996) and statistical-mechanics calculations (Lazaridis *et al.*, 1995). The results of these studies are often complementary to the experimental information already available and provide an important component to the current insight on solvent structure (McDowell & Kossiakoff, 1995) and function (Pomes & Roux, 1996; Oprea *et al.*, 1997). Furthermore, these techniques often provide the only means of obtaining an energetic understanding of some aspects of protein–water interaction.

The question of how the different techniques used to observe the location and properties of water molecules on the surface of proteins relate to and complement one another has been discussed in two short review articles (Levitt & Park, 1993; Karplus & Faerman, 1994). Karplus & Faerman discuss the reliability of each of the methods, illustrating their strengths and weaknesses, while Levitt & Park present the state of our understanding of protein–water interactions as it was in 1993, based on the synthesis of results obtained from the various methods discussed above.

23.5.3. Structural features of protein–water interactions derived from database analysis

The location and nature of water interaction with protein atoms are of great interest for understanding the role played by water molecules in the structural integrity and function of macromolecules. Baker & Hubbard (1984) presented an extensive analysis of hydrogen bonding in 15 proteins. A good portion of the study focused on hydrogen bonding with water. They observed that, in general, hydrogen bonds have a certain degree of flexibility, ranging in distance between 2.4 and 3.4 Å, with angular deviation from linear of up to 60°. The authors discussed the hydrogen-bonding geometry of water itself as well as the general aspects of the hydration of protein groups. Along the protein backbone, each carbonyl group is capable of making two hydrogen bonds, while amido groups make only one. Bifurcated hydrogen bonds are relatively rare, comprising only about 4% of the main-chain amido groups and even fewer of the side chains. Baker & Hubbard (1984) observed that of all of the hydrogen bonds made by water molecules, 42% are to main-chain carbonyl oxygens, 14% to main-chain amide groups and 44% to side-chain atoms. In a subsequent review that surveyed protein–water interactions, Savage & Wlodawer (1986) pointed out some of the major problems that hinder the accurate study of the precise hydrogen-bonding geometry and chemical features of protein–water interactions: the size of the biomolecular system, the resolution of the data, and the disorder of both the biomolecule and the solvent. The review was based on a comparison of X-ray and neutron diffraction studies of water interactions in a handful of proteins solved to a resolution of 1.5 Å or better with hydration properties in crystals of small- and medium-sized molecules solved to better than 1.0 Å resolution. Although a great deal had been learned about hydrogen-bonding properties of water in crystals of small molecules that presumably can be transferred to analogous interactions with protein atoms (Savage, 1986), the authors pointed out that for biomolecules there was, at the time, no consistent method being used for solvent analysis (Savage & Wlodawer, 1986). This problem was demonstrated and analysed in a subsequent review, where a comparison of three independently solved structures of interleukin-1 reveals a large variability in solvent structure (Karplus & Faerman, 1994).

The growing number of high-resolution protein crystal structures currently available in the Protein Data Bank (Berman *et al.*, 2000) allows for studies that extract statistically significant trends specific to protein–water interactions. The analysis of where and how water molecules bind to protein surfaces can be made at different levels. One can look at general properties of water interacting with each of the 20 amino-acid side chains, as well as with main-chain carbonyl oxygens and amido nitrogen atoms. At a higher level, one can study how these local interactions are modulated by the secondary-structure elements in which the residues are found. At the tertiary-structure level, one can study the location and function of water molecules as they are found in bridging secondary-structure elements and their role in the integrity of the protein architecture. At this level, studies regarding surface shape and hydrophilicity become important components of the analysis. Finally, the role of water molecules can be studied at the level of mediating protein–protein and protein–ligand interaction and their function in the affinity and specificity of these interactions. The remainder of this section summarizes information from database analysis of protein–water interactions at these various levels. The following sections then focus on individual examples to illustrate the classifications and functions of water–protein interactions.

23.5.3.1. Water distribution around the individual amino-acid residues in protein structures

The most comprehensive study of water molecules at the local level of binding to the individual types of amino-acid residues in protein structures was published in a series of papers (Thanki *et al.*, 1988, 1990, 1991; Walshaw & Goodfellow, 1993). The initial database consisted of 16 protein structures solved to better than 1.7 Å resolution and refined to an R factor of 26% or better (Thanki *et al.*, 1988). It was subsequently increased to 24 proteins using the same selection criteria (Thanki *et al.*, 1990, 1991; Walshaw & Goodfellow, 1993). All equivalent side chains as well as carbonyl or amide groups present in the database were brought to a common reference frame constructed from previously established bond lengths and bond angles (Momany *et al.*, 1975). The distribution of water molecules interacting with each of the 20 types of side chains was studied by focusing on particular atoms. Therefore, water molecules within 3.5 Å of N and O polar side-chain or main-chain atoms or within 5.0 Å of apolar side-chain carbon atoms were appropriately translated to the reference frame.

Fig. 23.5.3.1 shows the results of these superpositions for the polar main-chain amido and carbonyl groups as well as for some representative polar side chains: Ser, Tyr, Asp, Asn, Arg, His, Trp and Ala. The overall results show that despite the complex protein architecture, water molecules interact with hydroxyl, carbonyl and amide moieties, as well as with the sp^3-hybridized and ring nitrogen atoms, as expected from their known stereochemical requirements (Baker & Hubbard, 1984). Thus, there are water clusters in positions that optimize interaction with the lone-pair electrons on oxygen atoms and with the hydrogen atoms of amide and hydroxyl groups. Figs. 23.5.3.1(*a*) and (*b*) show the distribution of water molecules around the main-chain carbonyl oxygen and amido nitrogen atoms, respectively. The stereochemical requirements mentioned above are satisfied, with the distribution around the carbonyl oxygen clustered in two distinct regions peaking at an O–O distance of 2.7 Å. In contrast, there is

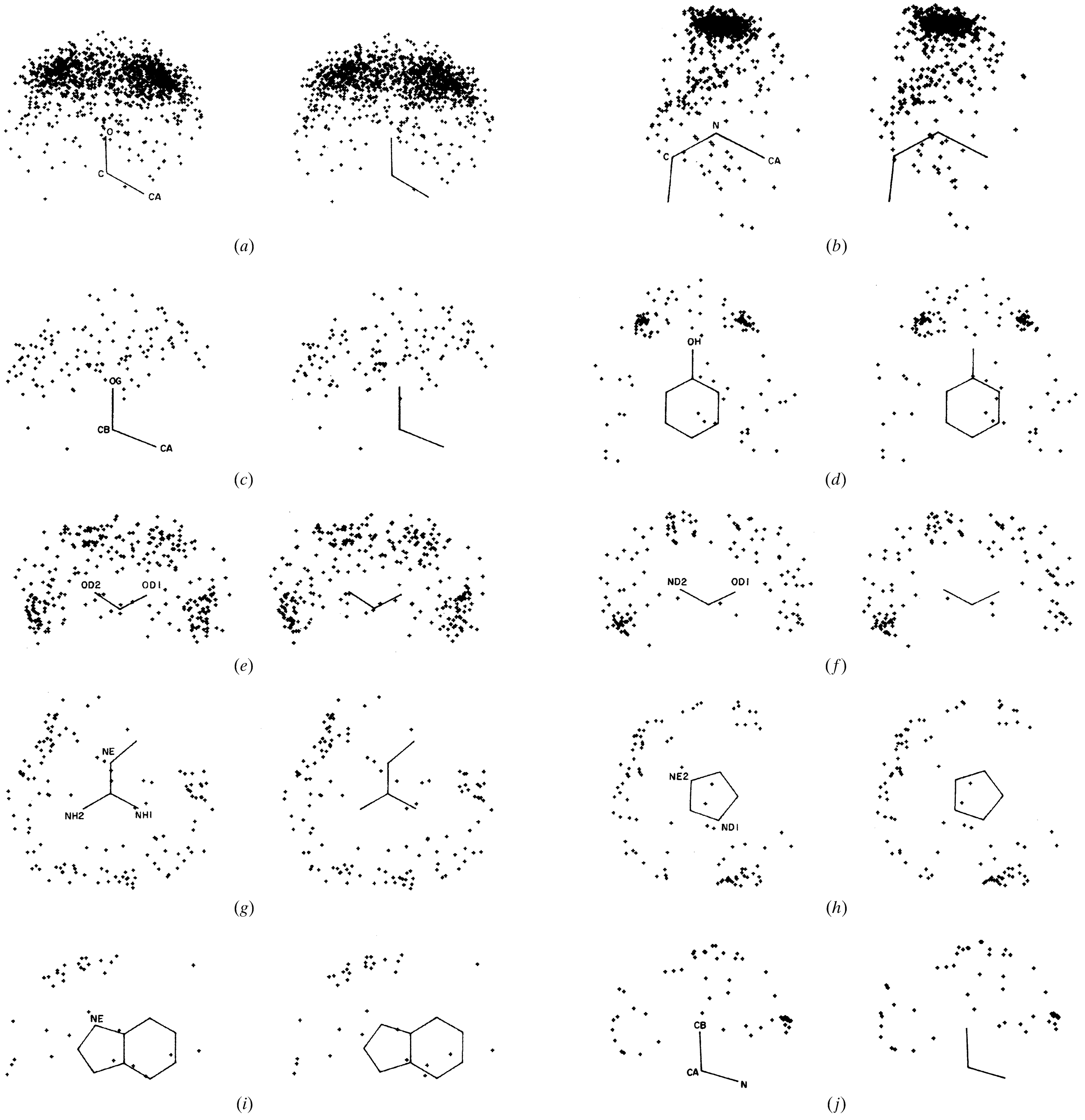


Figure 23.5.3.1

Distribution of water-molecule sites in stereo around: (*a*) main-chain O, (*b*) main-chain N, (*c*) Ser OG, (*d*) Tyr ring, (*e*) Asp OD1 and OD2, (*f*) Asn OD1 and ND2, (*g*) Arg NH1, NH2 and NE, (*h*) His ring to 3.5 Å, (*i*) Trp ring to 3.5 Å, (*j*) Ala CB. Reprinted with permission from Thanki *et al.* (1988). Copyright (1988) Academic Press.

a single water cluster interacting with the nitrogen, in line with the N—H bond at an N–O distance of about 2.9 Å. This cluster is much tighter than seen for the interactions with oxygen, reflecting a greater flexibility of water interaction with the carbonyl oxygen relative to the amido-group nitrogen atom.

Ser and Thr residues present a wide distribution of water molecules around the hydroxyl groups, presumably due to the freely rotating side chain. Fig. 23.5.3.1(*c*) shows the water-molecule distribution around Ser, which is only slightly different from that for Thr and can be representative of both. In contrast, the Tyr hydroxyl group is involved in resonance stabilization with the aromatic ring and, consequently, water molecules are clustered in the plane of the ring in well defined positions (Fig. 23.5.3.1*d*).

Fig. 23.5.3.1(*e*) shows the clustering of water molecules around the Asp side chain into four distinct groups, corresponding to the four available lone-pair electrons. The distribution around Glu is similar. Most water molecules interact with a single carbonyl oxygen, although about 11% (for Asp) and 15% (for Glu) of water molecules around these side chains interact with both oxygen atoms of a single carboxyl group. Water molecules that interact with Asn and Gln also show four clusters, with the two

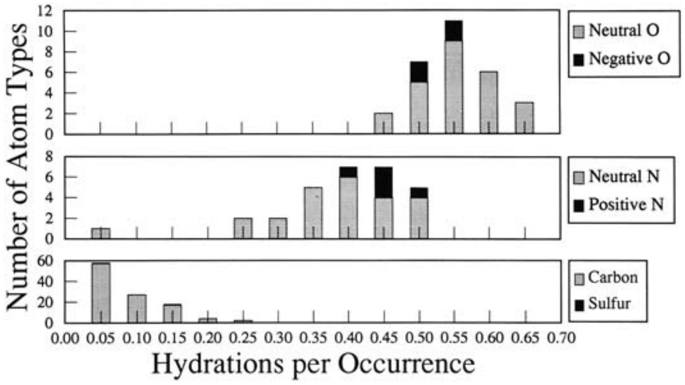


Figure 23.5.3.2

Distribution of atomic hydration values. To determine which atoms are similar or distinct with respect to water binding, we plotted the number of atom types (*e.g.* Ala amide nitrogen, Ala Cα, . . .) at each hydration per occurrence value. Each atom type contributed one vertical unit to the graph. Oxygen atoms were the most hydrated (top graph), with negatively charged oxygen (black bars) slightly less hydrated on average than neutral oxygen (grey bars). Nitrogens (middle graph) were the next most hydrated, overlapping the oxygen distribution, and positively charged nitrogens (black bars) were somewhat more hydrated than neutral nitrogens (grey bars). Proline's amide nitrogen, with no hydrogen-bonding capacity, had the lowest nitrogen hydration value (leftmost bar). Carbon and sulfur atoms (bottom graph; note change of *y*-axis scale) were the least hydrated, with sulfur values at 0.05 and 0.15 hydrations per occurrence. Reproduced from Kuhn *et al.* (1995). Copyright (1995) Wiley-Liss, Inc. Reprinted by permission of Wiley-Liss, Inc., a division of John Wiley & Sons, Inc.

clusters around the carbonyl group (C=O) less distinct than those around the amido (NH_2) group. Fig. 23.5.3.1(*f*) shows the distribution of water-molecule sites around Asn. In the case of Gln, the difference in water clustering around the carbonyl and amido groups is much less pronounced, possibly due to a greater degree of confusion in placing this longer side chain in the correct orientation. About 6% of the water molecules that interact with Asn or Gln are involved in hydrogen bonding to both the carbonyl oxygen and the amido nitrogen atoms.

The clustering of water molecules around the planar guanidyl group of Arg is distinctly positioned around the Nε atom and on either side of the NH1 and NH2 atoms. This is shown in Fig. 23.5.3.1(*g*). The clusters peak at a distance of about 3.0 Å from the nitrogen atoms. 7% of these water molecules are shared between NH1 and NH2, and only 3% are shared between the Nε and NH1 atoms. The distribution around the Lys side chain is much broader and is qualitatively similar to the one shown for Ser in Fig. 23.5.3.1(*c*), with no particular orientational preferences, mainly due to the freely rotating nature of the Cε—Nζ bond.

His and Trp are the two residues that contain ring nitrogen atoms, which comprise the main site of interaction with water molecules for these side chains. The distributions of water molecules within 3.5 Å of these residues are shown in Figs. 23.5.3.1(*h*) and (*i*). The clustering around His shows a peak at 2.7 Å and a larger peak at 3.1 Å. The closer peak corresponds to interactions with deprotonated nitrogen (Nδ), where the lone pair of electrons renders the deprotonated nitrogen more negatively charged than the corresponding protonated nitrogen (Nε) and, therefore, the deprotonated nitrogen pulls the water molecule closer. The peak at 3.1 Å is due to water interactions with the protonated nitrogen (Nε) of His. There is a strong preference for the water molecules to lie in the plane of the ring. Relatively few water molecules exist within 3.5 Å of Trp. They mostly cluster around the Nε nitrogen at varying distances. The number of water molecules interacting with His and Trp within 5.0 Å of the ring increases greatly and peaks at a distance of about 4 Å, as discussed below for hydrophobic residues in general (Walshaw & Goodfellow, 1993).

Table 23.5.3.1

Specific hydrophilicity values for protein atoms

Atom type	Hydrations per occurrence†
Neutral oxygen	0.53
Negative oxygen	0.51
Positive nitrogen	0.44
Neutral nitrogen	0.35
Carbon, sulfur	0.08

† The average number of hydrations per occurrence was calculated over all atoms within each group.

Overall, there seem to be weaker geometric constraints on oxygen acceptors compared to nitrogen donors. Furthermore, the water interaction with oxygen atoms peaks at a distance of about 2.8 Å, while the interactions with protonated nitrogen atoms occur at a somewhat longer distance of about 3.1 Å. This is possibly due to the larger van der Waals radius of nitrogen (1.8 Å) *versus* that of oxygen (1.7 Å) (Thanki *et al.*, 1988). A subsequent study of hydration around polar residues is based on seven proteins solved to better than 1.4 Å resolution (Roe & Teeter, 1993). The authors used cluster analysis to derive a predictive algorithm to locate water sites around polar side chains on protein surfaces, given the atomic coordinates of the protein alone. These more precise results confirm the general conclusions outlined above. The authors find that the water–oxygen distance is less than that of water–nitrogen by 0.07 Å and suggest the difference to be due to a van der Waals radius of 1.5 Å for nitrogen and 1.4 Å for oxygen (Roe & Teeter, 1993). Although the two groups cite different atomic radii for nitrogen and oxygen, this does not have an effect on the statistical analysis of the data. Roe & Teeter (1993) also find that the clusters associated with nitrogen atoms are approximately two times denser than those around oxygen atoms.

The analysis of the local water structure around the apolar side chains Ala, Val, Leu, Ile and Phe was extended to a distance 5.0 Å from the atom of interest, since these residues show only a few water molecules within the 3.5 Å cutoff used to analyse interactions with polar residues. The most noticeable observations from the analysis of apolar side chains are the water peak at a distance of 4 Å from the carbon atoms of interest and the presence of a polar protein atom within a hydrogen-bonding distance for 75% of these water molecules (Walshaw & Goodfellow, 1993). Phe prefers in-plane interactions and has peaks corresponding to the direction of the Cε1, Cε2, Cδ1 and Cδ2 atoms from the centre of the ring. Otherwise, any clustering observed for water molecules near apolar side chains is due to interactions with polar protein atoms and, consequently, is modulated by secondary structure.

A study of protein hydration based on atomic and residue hydrophilicity presents general results consistent with those discussed above, but also adds information that can be correlated with various experimentally and computationally derived hydrophilicity–hydrophobicity scales (Kuhn *et al.*, 1995). The authors used 10 837 water molecules found in 56 high-resolution protein crystal structures to obtain the average number of hydrations per occurrence over each amino-acid type and specific atom types. The hydration of the various amino-acid residues has already been discussed above. The atomic hydrophilicity values calculated for the different protein-atom types are of interest.

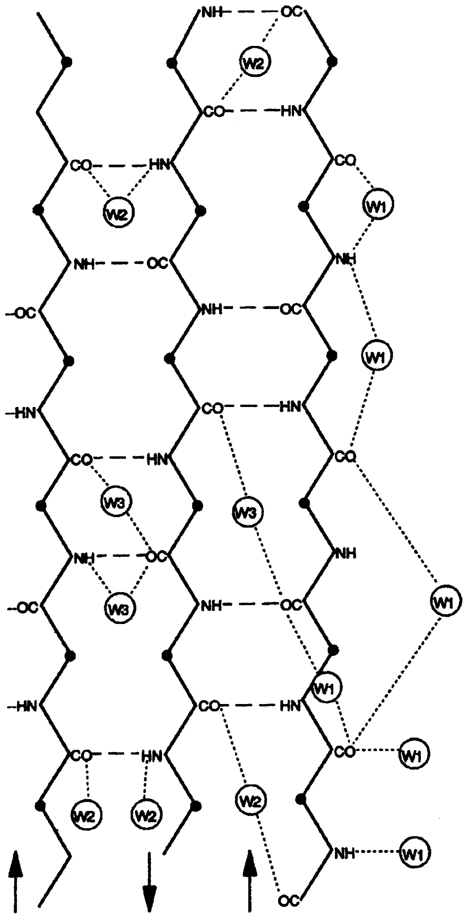


Figure 23.5.3.3
Diagram of edge (W1), end (W2) and middle (W3) categories of interactions of water molecules with main-chain atoms in antiparallel β-sheets. Reprinted with permission from Thanki *et al.* (1991). Copyright (1991) Academic Press.

Fig. 23.5.3.2 and Table 23.5.3.1 show that, regardless of where these atoms are found, neutral oxygen atoms exhibit the greatest hydration level per occurrence, closely followed by negatively charged oxygen atoms, which in turn are followed by positively charged nitrogens and neutral nitrogens, in that order. Carbon and sulfur atoms are indistinguishable in terms of hydration per occurrence and are grouped together as the least hydrated atoms (Kuhn *et al.*, 1995).

23.5.3.2. The effect of secondary structure on protein–water interactions

The main effect of secondary structure is on the hydration of main-chain carbonyl oxygens and amido nitrogen atoms. The clustering of water molecules around the small aliphatic apolar side chains (Walshaw & Goodfellow, 1993) and the Ser and Thr side chains (Thanki *et al.*, 1990) were also found to be guided by interactions with main-chain atoms belonging to a specific secondary structure. Other side chains are too large to have their hydration significantly affected by secondary structure. The broad solvent distribution around Ser and Thr side-chain hydroxyl oxygen atoms results from the combination of complex, but distinct, patterns that emerge when hydration around these side chains is examined separately in α-helices and β-sheets. Preferential hydrogen-bonding positions around Ser and Thr residues result from water molecules bridging between the hydroxyl group and another polar protein atom within the α-helix or β-sheet. These positions are dependent both on the χ_1 torsion angle and the type of secondary structure within which these residues are found (Thanki *et al.*, 1990).

The analysis of main-chain hydration focused separately on hydration of β-sheets, α-helices and turns (Thanki *et al.*, 1991). In general, more water molecules were found to interact with carbonyl oxygens than with amide groups, due primarily to the fact that carbonyl oxygen atoms can accept two hydrogen bonds, whereas amido groups can donate a single one. Thus, free carbonyl oxygen atoms have the potential to interact with two water molecules, whereas those already involved in a secondary-structure interaction with the protein still have a lone pair of electrons that can accept a hydrogen bond from a water molecule. Of the free carbonyl oxygen atoms within secondary-structure elements, 45% of those in α-helices and 68% of those in β-sheets interact with water molecules. Of those that are involved in secondary-structure interactions within the protein, 21% of those in α-helices and 17% of those in β-sheets also interact with solvent. The free amide groups are well hydrated, with 38% of those in α-helices and 54% of those in β-sheets interacting with water molecules. However, virtually none (2% in helices and 6% in sheets) of the amides already involved in secondary-structure hydrogen bonding also interact with a water molecule.

Three types of interactions were observed for water molecules in the context of β-sheets (Fig. 23.5.3.3). Most (68%) of these interactions are with the edge of the β-sheet, in an extension of the secondary structure. The second most prominent type of interaction, comprising 23% of the total, is at the ends of the β-strands with either free amido or carboxyl groups. Finally, only 10% of the water molecules are found to bridge between two strands in the middle of the β-sheet.

Interactions of water molecules with α-helices are also found in three distinct positions relative to the secondary structure (Fig. 23.5.3.4): at the carbonyl terminus of the helix, at the amido terminal end and in the middle. Of those interacting at the carbonyl terminus, 48% interact with the carbonyl oxygen alone, 11% also interact with a nearby main-chain atom and 41% are involved in a water-mediated C cap, bridging a small polar side chain (Ser, Thr, Asp, or Asn) to a free carbonyl group at the end of the helix. Of water molecules interacting at the amido terminus of the helix, 25% interact with free amido groups alone, 45% bridge to local main-chain atoms and many of the remaining mediate in N-cap interactions with small polar side chains such as Ser and Asp.

In general, turns have a high exposure to solvent and therefore are found to be well hydrated. The pattern of hydration varies both with the type of turn and the location of the atoms within the turn. Not surprisingly, there are about twice as many hydrogen bonds to carbonyl groups as there are to amide groups in turns. Although the majority of the water interactions with turns are to single carbonyl oxygen or amido nitrogen atoms, bridging water molecules do appear, especially within more open turns. They occur in a variety of different patterns, bridging between two main-chain atoms in the turn or between a main chain and a small polar side chain.

Clearly, water molecules play a functional role in maintaining the integrity of the secondary-structure elements of proteins. They are often seen to extend α-helices or β-sheets, serving as an interface between these secondary-structure elements and the bulk solvent. Water molecules are also found to mediate the interaction between two protein atoms within a given secondary structure that may be too far from each other to interact directly. This may be of great importance in turns, particularly the more open ones where the protein atoms are not in ideal positions to form a tight two-residue β-turn.

23.5.3.3. The effect of tertiary structure on protein–water interactions

At the tertiary level, there is an interdependence between protein surface shape and the extent of water binding (Kuhn *et al.*, 1992). Kuhn *et al.* (1992) studied the binding locations of 10 837 water molecules found in 56 high-resolution crystal structures using fractal atomic density and surface-accessibility algorithms. They found strong correlations between the positions of water molecules and protein surface shape and amino-acid residue type. A probe sphere with the radius of a water molecule revealed that, in general, protein surfaces exhibit convex groove areas and concave contact surfaces. Although grooves account for approximately one quarter of a given protein surface, they bind half the water molecules. Furthermore, only within grooves was hydration found to be dependent on residue type, with charged and polar residues as well as main-chain nitrogen and oxygen atoms exhibiting a greater degree of hydration than the non-polar residues. Outside the grooves, there was a low residue-independent hydration level, with no distinction between main-chain and side-chain atoms (Kuhn *et al.*, 1992). Levitt & Park (1993) discuss the paradox between the experimental observation that water molecules are crystallographically observed primarily in crevices (Kuhn *et al.*, 1992) and the results from theoretical calculations that argue that surface tension should make crevice waters bind less strongly (Nicholls *et al.*, 1991).

While the majority of the crystallographically observed water molecules appear on the outer protein surface, the internal protein packing is not perfect, so that the three-dimensional fold usually results in a number of internal cavities that can accommodate buried water molecules. The first analysis of such cavities was based on a small set of 12 proteins for which the authors characterized such sites by their size and area, as well as by whether or not they were occupied by crystallographically observed water molecules (Rashin *et al.*, 1986). More recently, two methodologically distinct studies of intramolecular cavities used much larger databases to provide extensive and mutually consistent conclusions regarding the properties of these sites (Hubbard *et al.*, 1994; Williams *et al.*, 1994). Hubbard *et al.* (1994) analysed 121 protein chains, with no two possessing a pairwise identity greater than 40%. This study is based on a systematic method of determining the shape as well as the size of the internal cavities and categorizes each cavity as either 'solvated' (with crystallographically visible water molecules) or 'empty' (with no crystallographically visible water molecules), noting the amino-acid-residue preferences in each type. Hydrogen-bonding patterns were also noted within the solvated sites. The second study (Williams *et al.*, 1994) selected 75 non-homologous monomeric proteins, solved at 2.5 Å resolution or better. Although the authors noted the general shape, size and location of cavities, the focus of this study was on the buried water molecules and the hydrogen-bonding patterns that they form within these sites.

In general, larger proteins are able to tolerate larger cavity sizes than small proteins, and nearly all proteins with more than 100 amino-acid residues are found to have at least one cavity. These cavities are found in the protein interior at a variety of distances from the surface and reflect the difficulty of perfect packing within the core. In the database of 121 proteins (Hubbard *et al.*, 1994), 265 cavities were found to be 'solvated' and 383 were 'empty'. The solvated cavities tend to be nearer to the protein surface than the empty cavities. Nearly 60% of the solvated cavities are occupied by a single water molecule and are of spherical shape. About 20% accommodate two water molecules, and 20% more are found to contain larger clusters (Williams *et al.*, 1994). These tend to have an elongated cigar shape. The cavity volume can be as large as 216 Å^3 (an elastase cavity containing seven water molecules). The solvated cavities tend to be larger than the empty ones, with average volumes of 39.4 and 20.7 Å^3, respectively (Hubbard *et al.*, 1994). The mean volume per water molecule in a cavity is 27 Å^3, as compared to 30 Å^3 in bulk water, suggesting that a water molecule is not favourably squeezed into a volume comparable to its own (11.5 Å^3), but rather occupies similar volumes upon transfer from the bulk into the protein interior.

Solvated cavities differ from empty ones not only in location and size within the protein, but also in the constitution of the amino-acid residues lining the cavity and the secondary-structure elements that are nearby. While 50% of the total cavity molecular surface is provided by polar atoms in solvated cavities, this fraction reflects only 16% of the empty cavity surface. Polarity, not size, is the predominant factor in determining the solvation state of a cavity. Interestingly, solvated cavities have more surface area provided by coil residues than the empty cavities, often found to be lined by residues in secondary structure (Hubbard *et al.*, 1994).

There is on average one buried water molecule per 27 amino-acid residues, although there is great variation

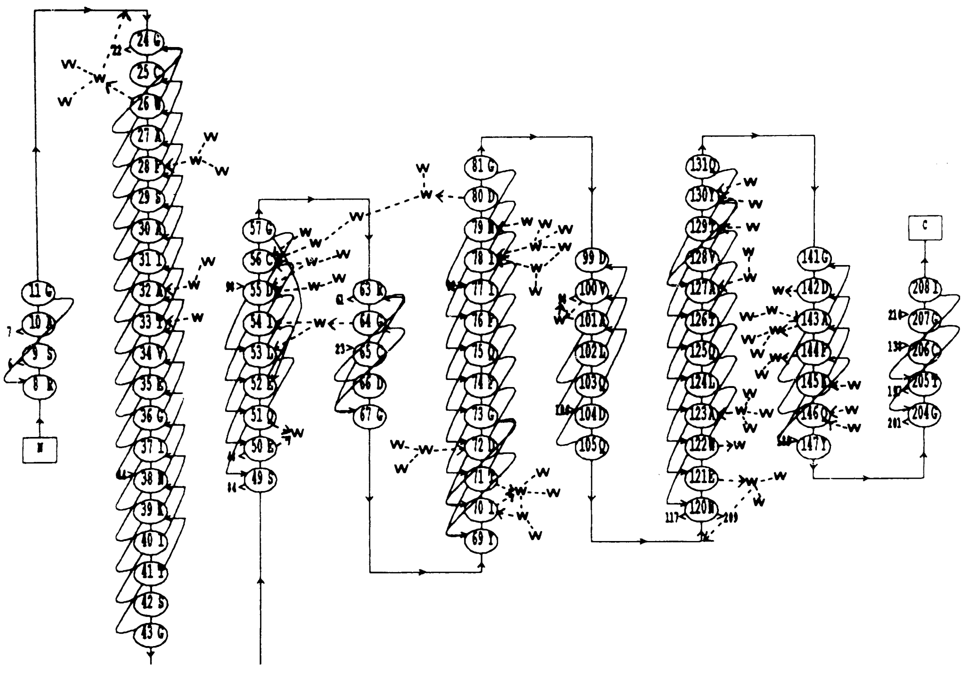
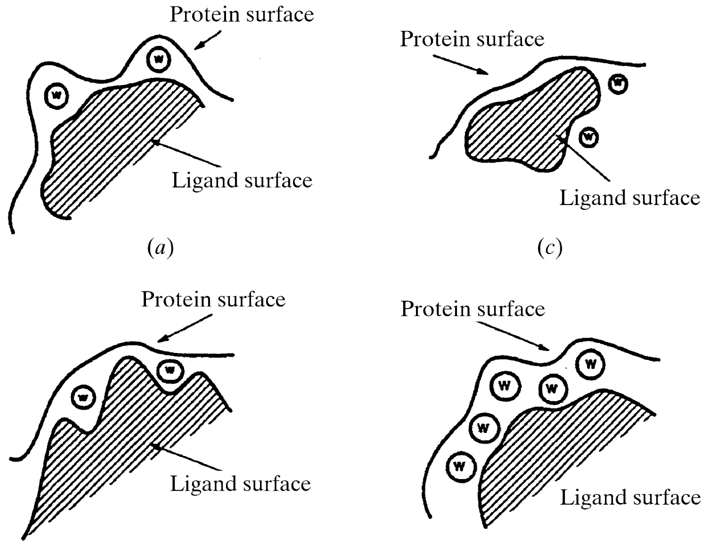


Figure 23.5.3.4
Diagram of the hydrogen bonds in the α-helical structure in actinidin. Reprinted with permission from Thanki *et al.* (1991). Copyright (1991) Academic Press.



(*a*) (*c*)

(*b*) (*d*)

Figure 23.5.3.5
Schematic illustration of water molecules bound in different types of grooves between protein and ligand. The hatched surfaces represent the ligand surface. (*a*) Water molecules bound in an indentation on the protein surface, where the protein surface area exposed to the water molecules is far larger than the ligand surface area; (*b*) water molecules bound in indentations on the ligand surface, where the ligand surface area exposed to the water molecule is larger than the protein surface area; (*c*) water molecules bound in shallow grooves at the protein–ligand interface and on the ligand surface; and (*d*) water molecules bound in clusters in elongated grooves with micro-grooves. Reprinted with permission from Poormina & Dean (1995*c*). Copyright (1995) Kluwer Academic Publishers.

between individual proteins. These water molecules most commonly form at least three hydrogen bonds with protein atoms or other buried water molecules. Only 18% of buried water molecules make two or fewer polar contacts. Of all of the hydrogen bonds made by buried water molecules, 53% are to protein backbone atoms, 30% to protein side-chain atoms, 17% to other buried water molecules, and 3% make no visible polar contacts at all (Williams *et al.*, 1994).

The appearance of cavities in the protein core is a consequence of the optimal packing of the protein polypeptide chain as it folds into the native, functional state. Where these cavities expose polar atoms to the hydrophobic protein core, one or more buried water molecules effectively become part of the structure, serving to maintain the protein integrity by fulfilling the hydrogen-bonding potential of atoms which are more favourably solvated.

23.5.3.4. Water mediation of protein–ligand interactions

A series of three papers presents the results of an analysis of water molecules mediating protein–ligand interactions in 19 crystal structures solved to better than 2.0 Å resolution and refined to an *R* factor of at least 23% (Poormina & Dean, 1995*a*,*b*,*c*). The studies focus on hydrogen-bonding features of water molecules bridging protein–ligand complexes (Poormina & Dean, 1995*b*), on the surface shape of the protein and ligand molecules at the water-binding sites (Poormina & Dean, 1995*c*), and on the structural and functional importance of water molecules conserved at the binding sites in five sets of evolutionarily related proteins (Poormina & Dean, 1995*a*). This study was largely motivated by an attempt to distinguish between properties of water-binding sites where water molecules are displaced by ligands and those where water molecules must be considered as part of the protein surface. This type of understanding has direct implications for drug and ligand design.

In general, there is a strong correlation between the number of water molecules found to bridge any given protein–ligand complex and the number of hydrophilic groups associated with the ligand. Within this context and in agreement with the conclusions of Kuhn *et al.* (1992), the authors found that the protein shape is important in determining the location of water-binding sites at the protein–ligand interface. Fig. 23.5.3.5 illustrates the different types of grooves observed in this study. Figs. 23.5.3.5(*a*) and (*b*) represent binding of bridging water molecules in deep grooves on the protein or on the ligand, respectively. The most common situation is illustrated in Fig. 23.5.3.5(*a*), with that in Fig. 23.5.3.5(*b*) occurring very rarely. Fig. 23.5.3.5(*c*) shows the situation where water molecules are found to interact with the ligand alone or at the periphery of the protein–ligand interface. Finally, Fig. 23.5.3.5(*d*) illustrates the situation where clusters of water molecules occupy elongated grooves, mediating the protein–ligand interaction. A striking example of this is given by the complex between chloramphenicol acetyl transferase and chloramphenicol, where two clusters of water molecules are found to form a layer between the enzyme and the ligand (Poormina & Dean, 1995*c*).

For the purposes of analysis, the authors distinguish between water molecules that interact with both protein and ligand, forming a bridge between the two, and water molecules that interact with either the protein or the ligand, but not with both. There is also a group of water molecules that interact with neither protein nor ligand, but are thought to contribute to the stability of the network of water molecules at the protein–ligand interface.

Of the 58 water molecules found to bridge between protein and ligand, 38 (nearly 80%) make three or more hydrogen bonds and satisfy tetrahedral geometry. Furthermore, they bind in deep grooves, generally interacting more strongly with the protein (Fig. 23.5.3.5*a*). The *B* factors of these bridging water molecules are comparable to those of the protein atoms with which they interact. They can, in effect, be considered an integral part of the protein structure and binding site. Many of these bridging water molecules are conserved throughout homologous proteins, even when different ligands are considered, and are clearly structurally significant in maintaining the properties of the protein binding sites.

Water molecules found to bind in shallow grooves do so either at the ligand surface or at the periphery of the protein–ligand interface. For many of these water molecules, the surface areas of the protein and the ligand exposed to the same water molecule are nearly equal. Water molecules binding in shallow grooves are found to have zero to two polar contacts with the protein and are not particularly well conserved within families of homologous proteins.

In general, the authors conclude that water molecules that are to be considered as part of the protein binding site during the design of a new ligand are those that bind in deep grooves, making multiple hydrogen bonds to protein atoms. These water molecules tend to be conserved through families of homologous proteins. The amino-acid residues that interact with deep-groove water molecules tend to be more conserved compared with other residues interacting with the ligand. Conversely, the binding of water in shallow grooves does not seem to be influenced by any special general feature of the protein or ligand surface, and it would be difficult to select water molecules *a priori* for inclusion as part of the protein structure during the process of ligand design.

23.5.4. Water structure in groups of well studied proteins

The analysis of general features of protein–water interactions derived from large databases provides an important context for the study of solvent structure in individual proteins. The number of crystallographically visible water molecules in any one X-ray structure depends on the resolution of the data, the degree of refinement of the model, the criteria used for placement of the less well defined water molecules, and on the experience of the crystallographer. Therefore, to differentiate between water molecules that have functional roles and those that associate randomly with the protein, it is desirable to determine commonalities between several independently solved structures of the protein of interest. There are different types of functional roles that can be determined at several levels. At the global level, one can find a small number of water molecules that are essential for the structural architecture common to a given family of homologous proteins. There are also those water molecules that are structurally important for a specific protein, being present in all independently solved structures of that protein, regardless of the crystal form in which the water molecule was determined or of its interactions with ligands. Water molecules that consistently appear in crystal structures of the protein solved in a specific space group but in no others may be important for crystal packing, but not to the integrity of the protein itself. Finally, a given water molecule may be essential for mediating in a protein–ligand complex, but never appear in the native protein. At this level, all of the independently solved structures of the complex would have the water molecule present. In the examples that follow, comparative analysis between carefully selected groups of structures reveals conserved water molecules at all of these different levels and shows how they carry out particular functional roles in specific examples.

23.5.4.1. Crystal structures of homologous proteins

There are two families of homologous proteins for which extensive solvent-structure comparisons have revealed water molecules important in maintaining structural features common to all members of the family. In the first study presented here, 35 crystal structures of eight members of the serine protease family were analysed (Sreenivasan & Axelsen, 1992), while the second study comprises a similar analysis of 11 independently solved structures of six members of the legume lectin family (Loris *et al.*, 1994).

23.5.4.1.1. Serine proteases of the trypsin family

The serine proteases have an especially large number of buried water molecules. Using a probe sphere of radius 1.4 Å, an iterative procedure was used to delete all accessible surface waters for each structure of chymotrypsin, chymotrypsinogen, trypsin, trypsinogen, elastase, kallikrein, rat tonin and rat mast cell protease. A total of 58 non-equivalent sites containing buried water molecules were found in the 35 crystal structures included in the study. Of these, 16 sites were common to all of the structures, with five additional sites common to proteins sharing the primary specificity of trypsin. A protein environment was defined for each of these 21 water sites to consist of the set of non-hydrogen protein atoms within 5 Å of the water oxygen atom. There are an average of 29 protein atoms per buried water molecule. Of these, 87% consist of main-chain atoms or conserved amino-acid side-chain atoms. The highly conserved nature of the amino-acid residues lining these water-binding sites suggests that the corresponding water molecules are important components of the protein tertiary structure and are likely to be present in all of the members of the trypsin family of serine proteases (Sreenivasan & Axelsen, 1992). Proteins in this family have two β-sheet domains, with the active site in the cleft between these domains. A large portion of the conserved buried water molecules occur in this cleft, mediating the interaction between the domains (Fig. 23.5.4.1). Conserved buried water molecules in other areas are found to bridge secondary-structure elements. These water molecules have been analysed extensively for elastase and are discussed in more detail below (Mattos, 2002).

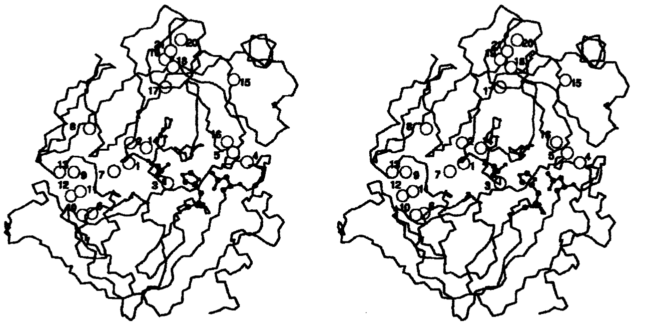


Figure 23.5.4.1
Stereoview of the set of 21 highly conserved buried waters in eukaryotic serine proteases. The trypsin backbone is represented as a stick drawing, with the catalytic triad at the centre (filled circles). Water molecules are represented as open circles. Reprinted with permission from Sreenivasan & Axelsen (1992). Copyright (1992) American Chemical Society.

23.5.4.1.2. Legume lectin family

Whereas the study on serine proteases described above focused on the buried water molecules, the study on the legume lectin family included all of the conserved water molecules in the first hydration sphere. A total of 11 crystal structures were superimposed, many of them containing two independently refined monomers, making a total of 21 crystallographically independent monomers (Loris *et al.*, 1994). The six different proteins in the family (lentil lectin, pea lectin, *Lythyrus* lectin, *Griffonia* isolectin IV, *Erythrina* lectin and concanavalin A) have sequence identities ranging from 100% to 40%. Water molecules in two superimposed crystal structures were considered to occupy the same site if they were within a predefined distance of 1 Å from each other. Seven water sites were found to be conserved in all of the family members included in the study. Four of these interact with the manganese and calcium ions, and one is in the ligand-binding site. The other two stabilize secondary structures: a β-hairpin turn and a β-bulge. In all cases, the protein composition of the site was strictly conserved. A larger number of water molecules are conserved within groups of closely related members of the family. The majority of these sites are found in the interface between the two monomers that come together to form a continuous 12-stranded β-pleated sheet and around the metal and monosaccharide binding regions (Fig. 23.5.4.2). Three crystal forms of lentil lectin were available for the study, and it was observed that of the 33 water molecules conserved between the corresponding three structures, none are involved in crystal contacts.

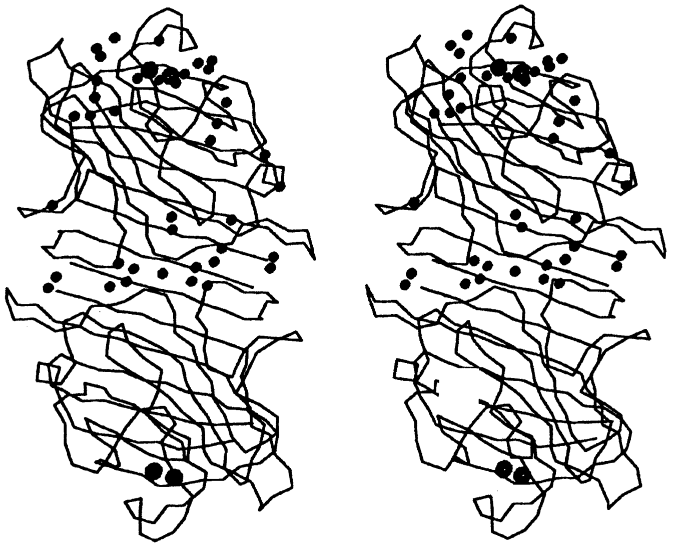


Figure 23.5.4.2
View of the 33 conserved hydration sites in the lentil lectin crystal structures superimposed on the backbone of the lentil lectin dimer. In order to emphasize the twofold symmetry, the waters at the dimer interface are shown for both lectin monomers. Reprinted with permission from Loris *et al.* (1994). Copyright (1994) The American Society for Biochemistry & Molecular Biology.

If one could generalize from the two studies described above, the conclusion would be that the water molecules strictly conserved across families of homologous proteins are found either at the binding site, at the interface between domains, or bridging secondary-structure elements which would otherwise not be part of the well defined protein architecture. Furthermore, it is clear that evolutionary pressure exists to maintain the composition of the amino-acid residues with which these crucial water molecules interact at their respective protein binding sites. A more recent study of conserved water molecules in a large family of microbial ribonucleases confirms the conclusions obtained in the two studies presented here (Loris *et al.*, 1999).

23.5.4.2. *Multiple crystal structures of the same protein*

Although not many studies have focused on the conserved water molecules across families of homologous proteins, there is currently a considerable amount of information on solvent structure based on groups of independently solved crystal structures of a specific protein. The comparison of multiple crystal structures is important to distinguish between the different roles played by water molecules on protein surfaces and to obtain a more complete picture of the first hydration sphere. In any one crystal structure of a given protein, it is extremely likely that the water molecules crucial to the structure or function of the protein will be seen in the electron-density map. However, the water molecules more loosely associated with the protein surface appear fortuitously in one or few structures, so that with every new structure one finds a series of water molecules not previously observed. A clear example of this is provided by a collection of eleven elastase structures solved in different organic solvents, where of a total of 1661 water molecules there are 178 molecules that are unique to one of the structures (Mattos & Ringe, 1996; Mattos, 2002; Mattos *et al.*, 2006).

23.5.4.2.1. *Elastase*

The crystal structure of porcine pancreatic elastase was solved in a variety of organic solvents, with the primary goal of mapping binding sites on the protein that could accommodate molecules representative of functional groups likely to be found in larger ligands (Ringe, 1995; Mattos & Ringe, 1996; Mattos *et al.*, 2006). Crystals of elastase cross-linked with glutaraldehyde were transferred to the following solutions: 100% acetonitrile, 95% acetone, 55% dimethylformamide, 80% ethanol, 40% trifluoroethanol, 80% isopropanol and 80% 5-hexene-1,2-diol (Mattos *et al.*, 2006; Mattos & Ringe, 1996). In general, the crystals did not diffract in most neat organic solvents. However, in the acetonitrile case, where they did, the result was striking. In the structure of elastase solved in >99% acetonitrile, there were 126 water molecules visible in the electron-density maps, indicating that a good portion of the first hydration shell of the protein was still present. In contrast, only nine molecules of acetonitrile were clearly identified in the electron-density maps (Allen *et al.*, 1996). This is a powerful assertion of the evolutionary specificity of water molecules for protein surfaces. Fig. 23.5.4.3 shows the clear contrast between the elongated electron density of an acetonitrile molecule and the spherical electron density of a water molecule.

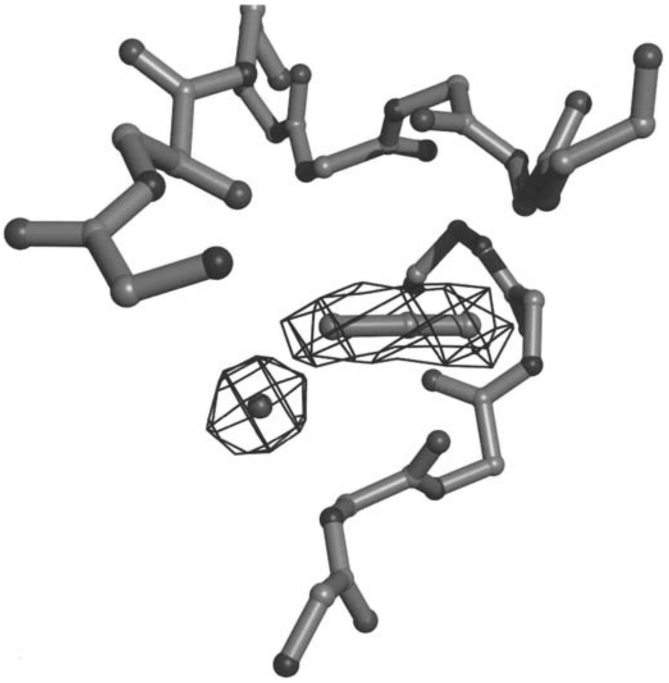


Figure 23.5.4.3
A $2F_o - F_c$ electron-density map contoured at the 1.2σ level shows a distinct ellipsoidal density for acetonitrile 707 and a spherical density for a nearby water molecule. The protein backbone of the binding pocket is represented with nitrogen atoms shown in dark grey, oxygen atoms in medium grey and carbons in a lighter grey. *MOLSCRIPT* (Kraulis, 1991) was used in the preparation of this figure. Reprinted with permission from Allen *et al.* (1996). Copyright (1996) American Chemical Society.

A similar result was obtained for all of the elastase structures solved in the mixtures of organic solvent and water mentioned above. A total of 11 structures were analysed, each containing 126–177 water molecules. The structures are listed in Table 23.5.4.1, together with the resolution of the data collected, the number of water molecules present and the number of organic solvent molecules observed in each case. The Cα superposition of the protein atoms in the 11 structures yielded a total of 1661 individual water molecules, occupying 426 unique water-binding sites on the elastase surface. Given that elastase has a total of 240 amino-acid residues, this represents a significant portion of the first hydration shell of the protein. This group of elastase structures served as a powerful source of information, leading to a

Table 23.5.4.1
Multiple-solvent crystal structures of elastase

Structure	Resolution (Å)	No. of water molecules	No. of organic solvent molecules
Cross-linked	1.9	165	0
Acetonitrile	2.2	126	9
Acetone	2.0	126	6
Dimethylformamide	2.0	153	6
Ethanol	2.0	135	12
Trifluoroethanol (1)	1.9	175	4
Trifluoroethanol (2)	1.85	177	3
Isopropanol	2.2	160	4
Benzene	1.9	162	4
Cyclohexane	1.95	135	7
5-Hexene-1,2-diol	2.2	147	5

classification of water types according to their interaction with the protein and an analysis of the specificity for water within each of the types determined (Mattos, 2002; Mattos *et al.*, 2006).

All of the 1661 water molecules were renumbered according to the site on the protein where they were found. Any two water molecules within 1 Å of a water molecule in the cross-linked elastase structure solved in distilled water (used as the reference structure) have a common number. Thirty-nine of the 426 water-binding sites were occupied in every one of the 11 structures and were considered structurally conserved. Among these are the 16 buried water-binding sites thought to be conserved among all serine proteases (Sreenivasan & Axelsen, 1992). The 26 remaining conserved water molecules are specific to elastase and are not necessarily buried. These water molecules in general tend to have low B factors, but a few have B factors in the 30–35 Å^2 range and one set of conserved water molecules have B factors in the 40 Å^2 range.

The classification of the water sites as buried, channel, crystal contact or surface was based on the number of hydrogen-bonding interactions that a water molecule at the site could make to the protein and involved no surface-accessibility calculations (Mattos, 2002). Water molecules were classified as buried if they made at least three good hydrogen-bonding interactions with protein main-chain atoms. A total of 23 buried water sites were identified in this manner, including 13 of the sites classified as buried by Sreenivasan & Axelsen (1992). One of the 16 serine protease conserved water-molecule sites is replaced by a His side chain in elastase (Sreenivasan & Axelsen, 1992). The remaining two serine protease conserved water sites were classified as channel based on the criteria used in the present study (see below). Interestingly, with the exception of these two channel water molecules, all of the buried sites found to be conserved in serine proteases are strictly conserved in all of the 11 structures in Table 23.5.4.1. The two channel water molecules are found in the aqueous structures of elastase, but are virtually absent in elastase transferred to organic solvents.

The water molecules occupying the 23 buried water sites identified in this study are tightly clustered when the protein Cα atoms are superimposed by least squares, and the interactions with the protein are conserved from structure to structure. Fig. 23.5.4.4 shows the positions of the buried water-binding sites in elastase. In general, they are found in the cleft between the two domains, in bridging elements of the secondary structure and at the base of water channels. This observation is consistent with the current understanding of the functional roles played by structurally conserved water molecules as discussed above and in the following sections.

The 29 water-binding sites classified as channel contain water molecules that make hydrogen bonds with at least two other water molecules within a protein groove. The analysis of a high-resolution crystal structure of elastase (1.65 Å) revealed seven channels with a total of 32 water-binding sites (Meyer *et al.*, 1988). All of these channels were also identified in the analysis of the 11 structures in Table 23.5.4.1. In addition, two other channels were observed. The locations of the nine elastase channels identified by the new criteria are shown in Fig. 23.5.4.5. Channels are often found in areas associated with buried water molecules, namely, at the crevice between the two domains and sandwiched between secondary-structure elements, where they lead from the surface of the protein to a buried water molecule. Fig. 23.5.4.5 also shows that the Cα superposition of the protein structures leads to a spread of water molecules within the channels. In any given structure, only two or three water molecules may be present, but the precise location and interaction with protein atoms vary so that when taken together the collection of structures gives a sense of flow inside the channels.

Of the remaining 374 water-molecule sites present within the 11 elastase structures included in this study, 56 were classified as crystal-contact sites and 318 as surface sites. Crystal-contact sites were considered to be occupied by water molecules that are within 4.0 Å of a symmetry-related protein molecule in the crystal. Fig. 23.5.4.6 shows the position of all the water molecules found to occupy these sites. The relatively large number of crystal-contact water-binding sites is a result of the somewhat broad criterion used to select them. Many of these sites are not

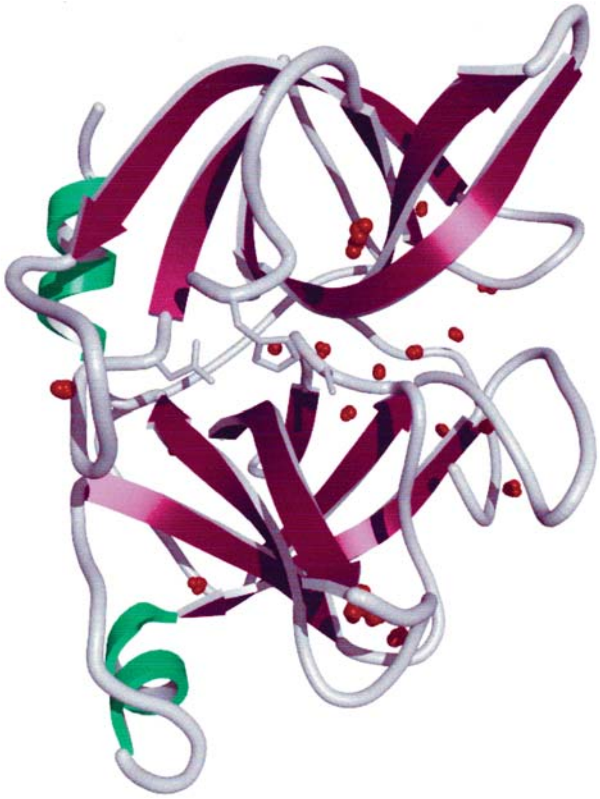


Figure 23.5.4.4
Crystal structure of porcine pancreatic elastase represented as a ribbon diagram using *MOLSCRIPT* (Kraulis, 1991). The two α-helices are shown in green, the β-sheets are in purple and the coils are in grey. Elastase contains 240 amino-acid residues, and is composed of two β-barrel domains. The catalytic triad (Asp108, His60 and Ser203) is shown explicitly. The buried crystallographic water molecules found in 11 superimposed elastase structures solved in a variety of solvents are shown in red.

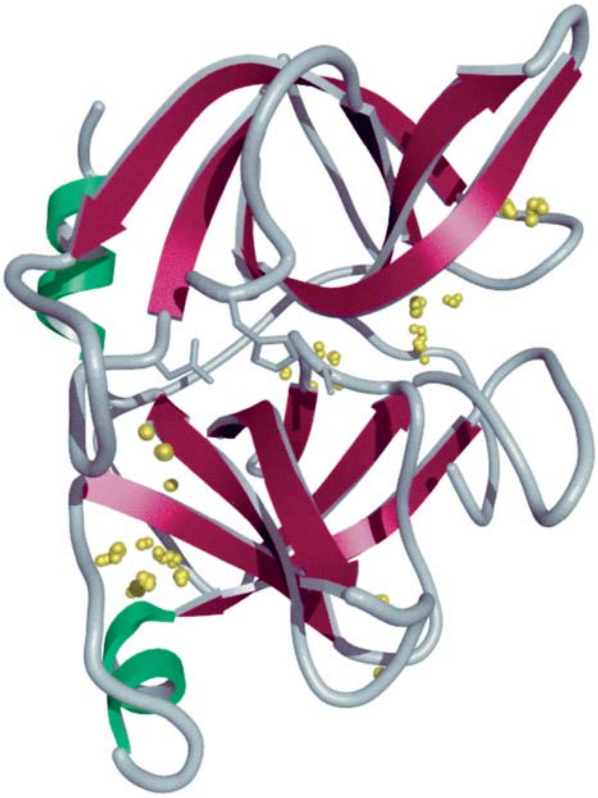


Figure 23.5.4.5
Elastase structure represented as in Fig. 23.5.4.4. The crystallographic water molecules found in channels in 11 superimposed elastase structures solved in a variety of solvents are shown in yellow.

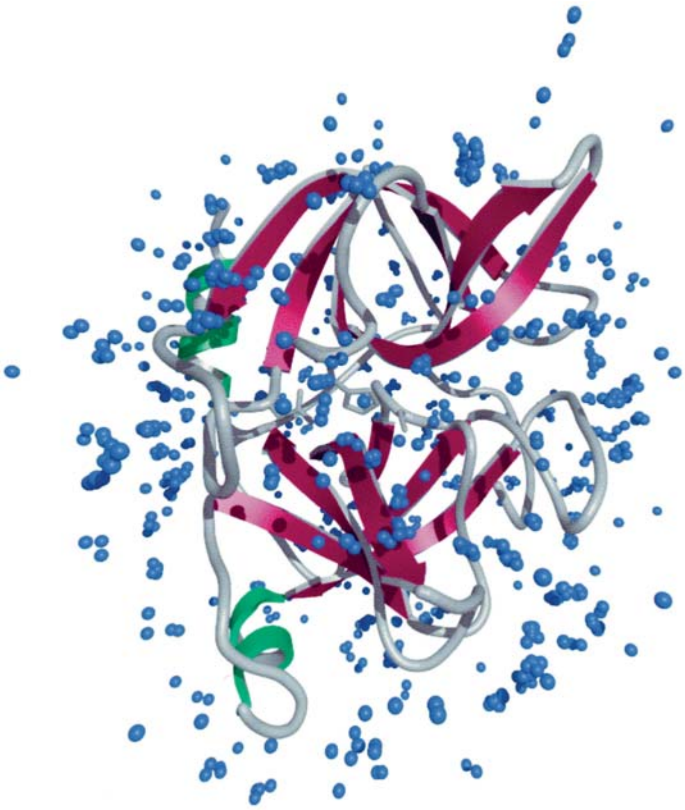


Figure 23.5.4.7
Elastase structure represented as in Fig. 23.5.4.4. The surface crystallographic water molecules found in 11 superimposed elastase structures solved in a variety of solvents are shown in blue.

within hydrogen-bonding distance from the nearby protein molecule, and most are not well conserved from structure to structure. Only eight of the 56 sites are occupied in the majority of the structures, and four of these make good multiple hydrogen bonds with two symmetry-related protein molecules in the crystal. These four water molecules seem to be structurally significant in the formation of the crystal contacts.

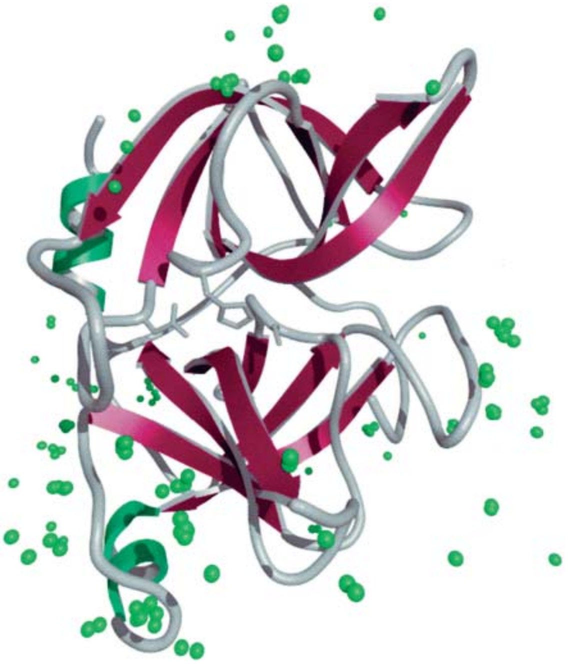


Figure 23.5.4.6
Elastase structure represented as in Fig. 23.5.4.4. The crystallographic water molecules involved in crystal contacts in 11 superimposed elastase structures solved in a variety of solvents are shown in green.

Surface water molecules were taken to be those that interact with side-chain protein atoms on the surface or make no more than two hydrogen bonds with backbone atoms. When the 11 structures are superimposed, the surface water molecules occupying a given site are not tightly clustered. Furthermore, there is flexibility in the interactions between these water molecules and the nearby protein atoms. For example, it is often the case that all water molecules within a surface site make two or three hydrogen bonds to protein atoms, but only one of them is conserved in all of the structures where the water molecule is present at the site. Fig. 23.5.4.7 illustrates the position of all of the surface water-binding sites. Although over half of these sites are occupied in at least two of the 11 structures, a good proportion of them (178) are found in only one of the structures considered.

While crystal-contact and surface water sites were classified separately, it is important to point out that, with the exception of the four crystal-contact water-binding sites mentioned above, the crystal-contact sites exhibit very much the same traits as the surface water sites. The difference is that in the latter case, the 'surface' is provided by a single protein molecule, while in the former the interaction between two symmetry-related protein molecules constitutes the surface with which the water molecules interact.

Of the 318 surface water molecules, 21 are in the active site. The active-site water molecules were selected to be those within 4 Å of any atom belonging to either the trifluoroacetyl-Lys-Phe-*p*-isopropylanilide (Mattos *et al.*, 1994) or the trifluoroacetyl-Lys-Pro-*p*-trifluoromethylanilide (Mattos *et al.*, 1995) inhibitors in the structures of their complexes with elastase. These inhibitors span a large area of the active site, including an exosite not occupied

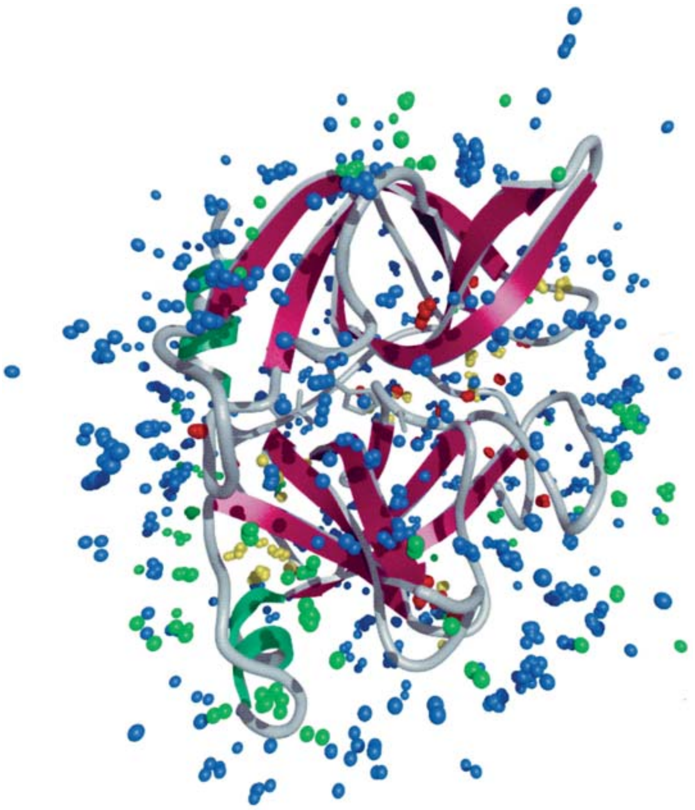


Figure 23.5.4.8
Elastase structure represented as in Fig. 23.5.4.4. The 1661 water molecules found in 11 superimposed elastase structures of elastase are colour-coded as in Figs. 23.5.4.4–23.5.4.7.

by substrate analogue inhibitors (Mattos *et al.*, 1994, 1995). The water-binding sites in the active site are not very well conserved, with most sites represented in only two to four of the 11 structures. When all of the structures are superimposed, there is at least one water molecule in each of the subsites in the elastase active site. These water molecules are displaced either by inhibitors or by organic solvent molecules in the various structures. It is not surprising that in elastase, a protein with relatively broad substrate specificity, the active site in the uncomplexed native protein is populated by many displaceable surface water molecules. With the exception of a water molecule present in the oxyanion hole, these water molecules tend to make a single hydrogen bond with the protein. This hydrogen-bonding interaction is not generally conserved between different structures where a given site is occupied in multiple structures. The displacement of these water molecules upon ligand binding is entropically favourable, as they are released into bulk solvent, without too much enthalpic cost. This relatively small enthalpic cost can be compensated by the protein–ligand interactions.

Fig. 23.5.4.8 shows all of the 1661 water molecules colour-coded by the various classifications described above. Clearly, the entire surface of the protein is well hydrated. Notice how the yellow channel waters are often followed by a red buried water molecule. In addition, there is often no obvious spatial distinction between molecules categorized as crystal contacts (green) and those categorized as surface (blue).

23.5.4.2.2. *T4 lysozyme*

Over 150 mutants of T4 lysozyme have been studied to date, and, for the majority of these, the crystal structures are available. Although most of the mutant structures crystallize isomorphously to the wild type, many of them provide a view of the molecule in different crystal environments. This collection of structures leads to the comparative analysis of the solvent positions in ten different crystal forms of T4 lysozyme, providing a clear picture of the effect of crystal contacts on the hydration sphere of a protein viewed by X-ray crystallography (Zhang & Matthews, 1994). The resolution and degree of refinement of the structures involved varied significantly, from 2.6 to 1.7 Å resolution, and the number of water molecules included per protein molecule ranged from 38 to 160. Nevertheless, this study revealed important features. A striking observation is that 95% of the solvent-exposed residues on T4 lysozyme were involved in at least one crystal contact in one or another of the crystal forms studied, showing that any part of the protein surface can be involved in crystal contacts. A corollary to this finding is that any of the surface water molecules can be displaced or involved in bridging protein–protein contacts in the crystal.

Of the 1675 individual water molecules observed in the 18 independently refined T4 lysozyme molecules included (Fig. 23.5.4.9), the ones that were within a sphere of radius 1.2 Å were considered to occupy the same site on the protein. As in the case of elastase described above, all of the water molecules observed upon superposition of the 18 T4 lysozyme structures represent a large portion of the first hydration shell. This reinforces the concept that multiple structures of a protein of interest provide a more complete picture of the protein hydration than possible with a single structure. There are four buried water sites that are occupied in at least 15 out of the 18 structures and are independent of crystal contacts. Two of these buried sites are at the hinge-bending region between the two helical domains and appear to play a functional role in the opening and closing of the

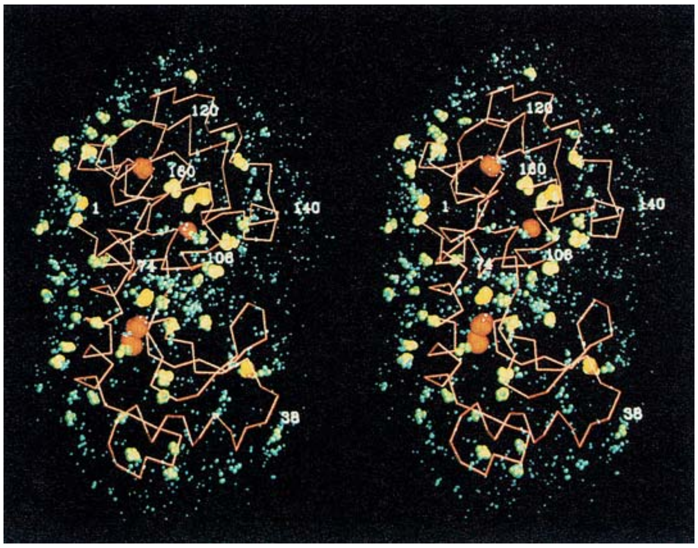


Figure 23.5.4.9
Distribution of solvent-binding sites in 18 mutant T4 lysozymes from ten refined crystal structures. The lysozyme structures were compared to identify common sites of hydration. A total of 1675 solvent molecules were included in the comparison. Each solvent molecule is represented by a coloured sphere. The size of the sphere is proportional to the number of lysozyme structures in which solvent was observed at the same site (*i.e.* within 1.2 Å). In addition, the colour of the solvent changes from blue for the least-conserved sites to red for the most-conserved ones [*e.g.* the red spheres indicate that solvent is observed with high frequency (15–17 times) at the four internal sites]. The numbers indicate representative residue positions along the backbone of the lysozyme molecule. Reprinted with the permission of Cambridge University Press from Zhang & Matthews (1994). Copyright (1994) The Protein Society.

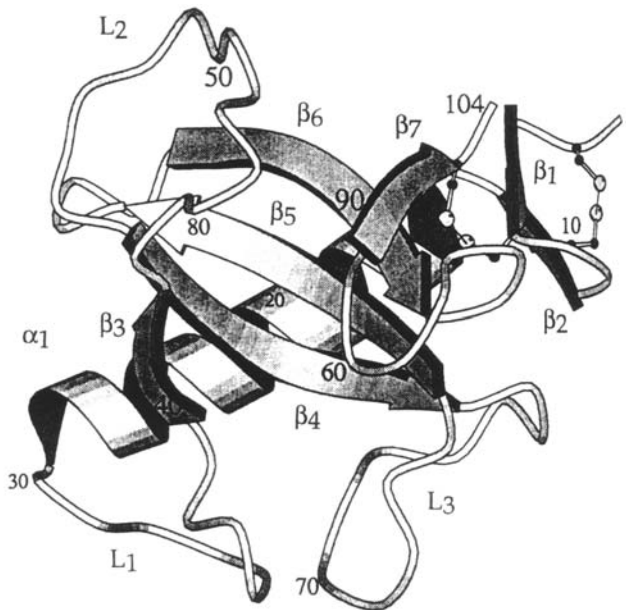


Figure 23.5.4.10
Three-dimensional structure of RNase T1. Secondary structure is denoted as follows: α_1, α-helix; β_n, strands of β-sheet structure; L_n, loops. Drawn using *MOLSCRIPT* (Kraulis, 1991). Residue numbers indicate the beginning and end of secondary-structure elements. Reprinted with permission from Pletinckx *et al.* (1994). Copyright (1994) American Chemical Society.

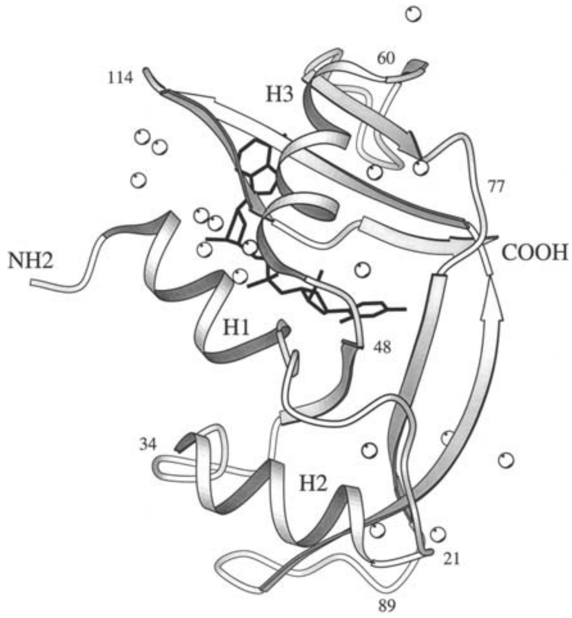


Figure 23.5.4.11
Overall structure of RNase A. The overall structure of the d(CpA) complex of RNase A is shown as a ribbon drawing using *MOLSCRIPT* (Kraulis, 1991). The conserved water molecules are shown as white spheres and the d(CpA) inhibitor in black. The three helices are labelled H1, H2 and H3. Reprinted with the permission of Cambridge University Press from Zegers *et al.* (1994). Copyright (1994) The Protein Society.

active site (Weaver & Matthews, 1987). The other two play a structural role at the protein core. Other than the four buried water molecules, the most conserved water sites appear at the active-site cleft between the two domains and at the N-termini of α-helices. As is the case in the previous works reviewed above, the 20 most conserved water sites appear in well conserved protein environments and generally have low temperature factors. Buried or highly conserved water molecules also tend to make at least three hydrogen bonds with protein atoms or other water molecules. The less-conserved water sites appear more randomly on the protein surface and are strongly influenced by the particular crystal environment in which the structure was solved.

23.5.4.2.3. Ribonuclease T1

A group of four crystal structures of ribonuclease T1 in complex with guanosine, guanosine-2′-phosphate, guanylyl-2′,5′-guanosine and vanadate were used for an analysis of conserved water positions that contribute to the structural stabilization of the protein (Malin *et al.*, 1991). The four structures were obtained from isomorphous crystals and ranged in resolution from 1.7 to 1.9 Å. Conserved water molecules were considered to be those found within a sphere of 1 Å from each other in all four structures. All other water molecules were excluded from the analysis. 30 water molecules were found to be conserved. Of these, ten were observed near crystal contacts, although only one appears to be dictated by the crystal contact itself, making a single hydrogen bond with each of the symmetry-related protein molecules. Ten other water molecules form a channel that brings together an α-helix and a hairpin-like loop structure and then go on to wrap around the calcium ion, providing half of its coordination sphere. The first five of these water molecules are completely buried, holding together the two secondary-structure elements, which would otherwise collapse (Malin *et al.*, 1991). Two water molecules are found to stabilize the N and C termini, which are brought together by a disulfide bond. The remaining eight conserved water molecules hold together various elements of secondary structure or are located in the active site.

An interesting extension to this study included four additional structures: the E58A mutant in complex with guanosine-2′-monophosphate, the H92A mutant crystallized under two different conditions and wild-type RNase T1 in complex with guanosine-3′,5′-biphosphate. Two of these crystal forms were not isomorphous with the native protein crystals or with each other. Thus a total of eight structures solved in three different space groups were analysed (Pletinckx *et al.*, 1994). Although the effect of crystal packing on the three-dimensional structure of the protein is minimal, there are some significant differences in the solvent structure. In particular, there is no evidence of the calcium-binding site and its coordinating water structure in any crystal forms other than the canonical wild type. Instead, the E58A mutant has a sodium-binding site at a different position, along with three previously unobserved water molecules. It is clear that the presence of the metal ions is fortuitous and linked to the crystallization conditions.

There are 25 water molecules structurally conserved throughout the different packing arrangements studied. Ten of these are single sites, there are three clusters of two water molecules and a larger cluster originally described by Malin *et al.* (1991) to hold together the core of the protein. As was observed for the study on T4 lysozyme (Zhang & Matthews, 1994), the strictly conserved water-binding sites present in crystal structures solved across different space groups are involved in bridging protein secondary-structure elements and seem to be crucial for the integrity of the protein structure.

23.5.4.2.4. Ribonuclease A

Ribonuclease A is not homologous to ribonuclease T1 in either sequence or structure, but both have evolved to catalyse the same reaction with specificity for different substrates (compare Figs. 23.5.4.10 and 23.5.4.11). Ribonuclease A cleaves RNA after pyrimidines, while ribonuclease T1 cleaves specifically after guanine. Therefore, the information obtained from a study of the solvent structure in ribonuclease A is completely independent from that described above for ribonuclease T1. A collection of ten crystal structures of ribonuclease A, derived from five different crystal forms, were compared pairwise after least-squares superposition (Zegers *et al.*, 1994). Seventeen conserved water molecules were found to be within a sphere of 0.5 Å of each other in all of the ten structures and are shown in Fig. 23.5.4.11. These water molecules were found in small clusters of two or three or as part of a larger solvent network. Not surprisingly, they form multiple hydrogen bonds with the protein and generally have low temperature factors. Of the 17 structurally conserved sites, 13 are associated with one of the three α-helices. Most of these link the helices to one of the β-strands. Three water molecules are involved in hydrogen bonding with unpaired amido and carbonyl groups on the protein, and one is found on top of the β-pleated sheet. These interactions result in bringing together elements of secondary structure and in stabilizing distortions within these elements. Conserved water molecules are also responsible for bridging the N-terminal helix to the C-terminal β-strand, which form the two halves of the active site. Conserved water molecules in the active site of ribonuclease A have more recently been analysed using multiple solvent crystal structures (Dechene *et al.*, 2009).

23.5.4.2.5. Protein kinase A

The comparative study of water molecules in seven different protein kinase A structures in complex with different ligands focused exclusively on the active site (Shaltiel *et al.*, 1998). All of the structures were solved from isomorphous crystals to resolutions ranging from 2.0 to 2.9 Å. The more lenient cutoff of 1.5 Å for the radius of the sphere within which the conserved water molecules must be found among the different structures is consistent with the relatively low resolutions of the structures included in this study. The group of structures represents the open, the closed and an intermediate conformation of the catalytic kinase domain. There is a set of six conserved water sites in the active site, in addition to the ATP molecule and the magnesium ion. The conserved water molecules coordinate to ATP, the metal ion and a conserved Tyr residue from the carboxyl terminus of the protein. Thus, the active site consists of an extended network of interactions that weave together both domains of the core, with water molecules playing an integral role in maintaining the structural features important for catalysis. Many of these water molecules associate directly with the inhibitors. In addition, five water sites are observed in positions that would be occupied by substrates or substrate analogues. These water molecules are displaced by ligand oxygen atoms that can compensate for the water hydrogen-bonding interaction with the protein.

23.5.4.3. Summary

Water molecules associated with proteins can be divided between those that are conserved as a result of their functional significance and those that are partially conserved or not conserved at all. The conserved water molecules are generally classified as buried or channel (by a variety of criteria). They tend to be present in the clefts between domains, are critical components of active sites, or bridge between secondary-structure elements. The water molecules that are not conserved occupy hydration sites with favourable hydrogen-bonding characteristics, where the presence of a water molecule is not essential for the structural or functional integrity of the protein.

The displacement of water molecules by organic solvent molecules in the elastase work described above showed that most displaced waters are those classified as surface or crystal-contact waters (Mattos *et al.*, 2006). In the three cases where a buried water molecule was displaced, an alcohol hydroxyl oxygen was found to replace the protein–water hydrogen-bonding interactions. This is analogous to the active-site water molecule in the HIV aspartate protease that gets replaced by a carbonyl group of a potent cyclic urea inhibitor (Lam *et al.*, 1994). In these situations, release of a tightly bound water molecule is entropically favourable, and its enthalpic interactions with the protein are compensated by similar protein–ligand interactions.

The effect of crystal contacts on the water structure was clearly illustrated in the T4 lysozyme work (Zhang & Matthews, 1994). The internal structurally conserved water molecules are unaffected by crystal contacts. Conversely, any of the surface water sites are potentially available either to be replaced by or to mediate crystal contacts, as 95% of the T4 lysozyme surface is involved in a crystal contact when all ten crystal forms are taken together.

23.5.5. The classic models: small proteins with high-resolution crystal structures

Crambin and BPTI are among the handful of proteins for which X-ray crystal structures have been obtained to 1 Å resolution or better. In general, these proteins are relatively small (BPTI, the largest in this group, has 58 amino-acid residues) and often contain at least one disulfide bond. These high-resolution crystal structures have provided structural information beyond that available for larger proteins, particularly with respect to the surface solvent structure. The available detail and precision of the structures, as well as their small size, make them ideal models in computational studies of protein energetics and dynamics. Both crambin and BPTI were used during the pioneering years of protein molecular-dynamics calculations. In this section, special attention is given to crambin and BPTI as representative proteins for which very high resolution structures are available. Focus is on the features of solvent structure that are not available for other proteins.

23.5.5.1. Crambin

Crambin is a plant-seed hydrophobic protein of unknown function. It contains 46 amino-acid residues and was reported to form crystals that diffract to 0.88 Å resolution (Teeter & Hendrickson, 1979). The crystal structure of crambin was determined to 0.945 Å resolution directly from anomalous scattering by the six sulfur atoms involved in three disulfide bonds (Hendrickson & Teeter, 1981). Crambin is an amphipathic molecule in that the hydrophilic components (including six charged groups) are segregated from a mainly hydrophobic surface.

A total of 64 water molecules and two ethanol molecules were located in the electron-density map, despite the fact that the structure was determined in 60% ethanol. The overwhelming number of water molecules compared to ethanol is consistent with the results of the multiple-solvent crystal structures experiments described above for elastase (Mattos *et al.*, 2006).

Most of the 64 water molecules found in crambin interact with polar side chains in the typical manner described previously. The unusual information about solvent structure offered by the

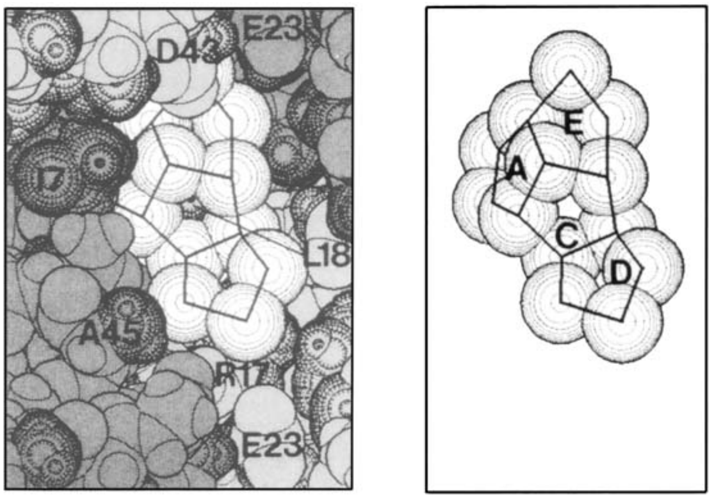


Figure 23.5.5.1
van der Waals surface diagram of the water pentagons A, C, D and E in crambin viewed in the negative **a** direction. Rings A, C and E form a cap around leucine 18. Hydrophobic atoms are shown as dark circles, and water oxygens are shown as light circles. The methyl group of leucine 18 can be seen through the C ring. Adjacent translationally related molecules are shaded. The van der Waals radii used for the protein C, N and O atoms are 1.7, 1.4 and 1.4 Å, respectively, and for water oxygen, 1.8 Å. The larger radius is used for the water oxygens because hydrogen atoms have been omitted. Reprinted with the permission of the author from Teeter (1984).

crambin model is that the arrangement of water molecules around hydrophobic residues is similar to that observed for clathrate hydrate structures (Teeter, 1991). Pentagonal water rings are observed to cap the Cδ2 atom of Leu18 as well as the hydrophobic methylene groups of Arg17 (Teeter, 1984, 1991). The set of five connected water rings is shown in Fig. 23.5.5.1. This ring cluster extends toward the protein, forming heterocyclic rings that are described in detail in the original article (Teeter, 1984).

Although crambin provides the clearest example of pentagonal water rings on a hydrophobic protein surface, it is not the only one. Other high-resolution crystal structures (better than 1.4 Å), such as insulin and cytochrome *c*, have also revealed pentagonal rings, but never to the extent seen in crambin (Teeter, 1984). This is very likely to be a general mode of interaction between water and hydrophobic moieties, be it in inorganic, organic, or biological molecules. The fact that it is not observed in protein structures in general may be related to the lower resolution of most X-ray structures, where it is not possible to model the more disordered areas where these patterns are likely to be found.

23.5.5.2. Bovine pancreatic trypsin inhibitor

Bovine pancreatic trypsin inhibitor (BPTI) is a protein of 58 amino-acid residues whose X-ray crystal structure was obtained in the original crystal form to 1.5 Å resolution (Deisenhofer & Steigemann, 1975). Subsequently, 1 Å X-ray data were obtained from a different crystal form, and the new model was jointly refined with 1.8 Å neutron diffraction data (Wlodawer *et al.*, 1984). Minor differences in structure between the two crystal forms of BPTI were observed (Wlodawer *et al.*, 1984). The interesting contribution of the 1 Å model to the understanding of solvent structure resulted from the ability to refine occupancy at this resolution. A total of 63 water molecules were placed in the model, 20 of them within 1 Å of a water molecule found in the structure solved in the original crystal form. During refinement against the 1 Å data set, full occupancy was assigned to all protein atoms, and water occupancy was allowed to refine. Of the 63 water-molecule positions, 29 were found to be fully occupied. The remaining 34 had partial occupancies, with 0.4 being the minimum occupancy found. Given that there are very few contacts between protein molecules in the crystal (Wlodawer *et al.*, 1984), it is reasonable to assume that this observation is representative of water occupancies on protein surfaces in general. It is likely that well over half of the water positions found on protein surfaces are less than fully occupied, although there is no definitive proof that this is true.

23.5.5.3. Summary

In general, small proteins serve as important models through which more precise details of protein–water interactions can be obtained due to the very high resolution to which their structures can be solved. With respect to understanding solvent structure, the two major contributions of these very high resolution protein models are the observation of solvent structure around hydrophobic residues, where at lower resolution the water molecules 'look' disordered, and a glimpse at the pattern of water occupancy likely to occur on protein surfaces.

23.5.6. Water molecules as mediators of complex formation

The examples given in Section 23.5.4 illustrate two important roles played by water molecules at the binding sites of proteins: as structural water molecules and as displaceable water molecules. As a structural part of a binding site, water molecules are found to be strictly conserved. They are either involved in stabilizing the coming together of secondary-structure elements in a way that appropriately shapes the binding site, or they fill grooves on the protein surface, making it more specific for a given ligand. The second role involves the presence of less tightly bound, partially conserved water molecules that get displaced by the ligand upon binding. In the few examples where tightly bound water molecules are displaced by a ligand, the hydrogen-bonding interaction of the water with the protein is replaced by an atom on the ligand. A third role, not yet discussed, of water molecules in protein active sites is in the catalytic mechanism of enzymatic reactions. An extensive network of water molecules near the active site of serine proteases has been implicated in the catalytic mechanism of these enzymes (Meyer *et al.*, 1988; Meyer, 1992). If this hypothesis is indeed correct, it provides a good example of the cooperation between water molecules and protein atoms in the optimization of function. Unfortunately, it is difficult to explicitly detect catalytic water molecules crystallographically, due to the long data-collection time relative to a catalytic event. However, the development of time-resolved Laue diffraction methods has provided a view of the catalytic water molecule in some proteins, *e.g.* trypsin (Singer *et al.*, 1993), and progress is likely to continue in this area. This section focuses on a few particular examples of how water molecules mediate the formation of complexes, either in the active sites of enzymes or in the binding interface between macromolecules or protein–ligand complexes.

23.5.6.1. Antigen–antibody association

The X-ray crystal structures of the Fv fragment of the monoclonal antibody D1.3 and the structure of its complex with hen egg-white lysozyme were both solved to 1.8 Å resolution (Bhat *et al.*, 1994). This study revealed a significant number of water molecules contributing to the chemical complementarity at the antigen–antibody interface. There are 23 water molecules at the antigen-binding site of the free antibody fragment, while 48 are present mediating complex formation. Seven water molecules are in equivalent positions in the free and complexed antibody (within 1.5 Å). There is no net loss of water molecules at the combining site. In fact, the total number of water molecules at the antigen–antibody interface is not less, but more, than the sum of those in the free antibody combining site and in the antigenic determinant. Furthermore, there is a general decrease in *B* factors of the binding-site residues upon complex formation, implying a decrease in entropy (Bhat *et al.*, 1994). The structural results indicate that water molecules at the antigen–antibody interface play a variety of important roles. Some form an integral part of the binding site, fine-tuning the shape and charge complementarity of the interaction. Others are found to be displaced during complex formation, and still others are unique to the complex, bridging between the two molecules in a variety of locations throughout the complex interface.

The structural analysis correlated well with results of calorimetric experiments that showed that complex formation is enthalpically driven, with an unfavourable entropic contribution (Bhat *et al.*, 1994). The authors suggest that water molecules play a central role in mediating complex formation and claim that the hydrophobic effect is not important in this case. This is an argument that goes contrary to the idea that affinity is contributed by hydrophobic interactions within a relatively small portion of the interface between the interacting molecules, with hydrogen-bonding and charge–charge interactions contributing primarily to specificity (Hendsch & Tidor, 1994; Clackson & Wells, 1995; Hendsch *et al.*, 1996).

23.5.6.2. Protein–DNA recognition

The *trp* repressor binds specifically to the target DNA sequence ACTAGT, resulting in the transcriptional control of L-tryptophan levels in bacteria. The crystal structure of the *trp* repressor/operator complex was solved to 1.9 Å resolution (Otwinowski *et al.*, 1988). Although the structure revealed hydrogen-bonding interactions between the protein and the backbone phosphate groups, no direct hydrogen bonds or non-polar contacts between the protein and DNA bases were observed. Specificity was therefore attributed to the effect of the sequence on the geometry of the phosphate backbone and to water-mediated polar contacts between protein atoms and specific DNA bases. To confirm this hypothesis, the 1.95 Å resolution crystal structure of the free decamer CCACTAGTGG was obtained, containing the recognition six-base-pair sequence (Shakked *et al.*, 1994). A comparative analysis of the free and complexed DNA showed that, when bound to the *trp* repressor, the six-base-pair region is bent by about 15° so as to compress the major groove, with concomitant expansion of the minor groove relative to the uncomplexed DNA (Shakked *et al.*, 1994). However, both free and complexed DNA are underwound, with 10.6 base pairs per turn, rather than the usual 10.0 base pairs per turn. This feature is presumably a result of the particular DNA sequence and is thought to decrease the energy barrier for the binding interaction with the *trp* repressor protein (Shakked *et al.*, 1994). Another specificity component suggested by the authors is conferred by the hydration of the consensus bases. Ten water molecules are observed to interact in the major groove at similar positions in both the free and complexed DNA. Three of these mediate in four hydrogen-bonding interactions to the protein in the complex. Interestingly, the DNA bases to which these three water molecules are bound are among the most conserved and mutationally sensitive bases of the operator. In effect, these three water molecules can be regarded as extensions of the DNA bases and part of the specific recognition elements of the target DNA sequence (Shakked *et al.*, 1994).

The idea of water molecules as mediators of interactions conferring specificity in protein–DNA associations is further supported by the co-crystal structure of the HNF-3/fork head DNA-recognition motif in complex with DNA, solved to 2.5 Å resolution (Clark *et al.*, 1993). Although the lower resolution of this protein–DNA complex may limit the unambiguous determination of water molecules to those that are tightly bound, a series of water molecules are observed in the major groove, bridging specific DNA bases to amino-acid side chains in one of the α-helices of the protein. In this case, direct hydrogen bonding between DNA bases and protein side chains also exists.

The involvement of water in specific protein–DNA recognition was further confirmed in a study of the accuracy of specific DNA cleavage by the restriction endonuclease *Eco*RI under different osmotic pressures (Robinson & Siglar, 1993). Changes in osmotic pressure, resulting from changes in osmolite concentrations, have direct effects on the number of water molecules associated with macromolecules (Rand, 1992). The *Eco*RI experiments show that water activity affects site-specific DNA recognition, with an increase in osmotic pressure leading to a decrease in accuracy of protein–DNA recognition, as observed by DNA cleavage at sites containing an incorrect base pair (Robinson & Siglar, 1993). The results of this study strongly imply a role for one or more water molecules in recognition of specific sequences of DNA. The authors suggest that water mediation may constitute a general motif for sequence-specific DNA recognition by DNA-binding proteins (Robinson & Siglar, 1993).

The role of water molecules as mediators of sequence-specific DNA recognition may be a general motif, but not a necessary one. The solution NMR structure of the complex of erythroid transcription factor GATA-1 with the 16-base-pair DNA fragment GTTGCAGATAAACATT, containing the recognition sequence, shows that the specific interactions between GATA-1 and the major groove of the DNA are dominated by van der Waals interactions hydrophobic in nature (Omichinski *et al.*, 1993). Furthermore, NMR experiments designed to identify the location of water molecules in the complex detected clusters of water molecules bridging the protein to the DNA phosphate backbone, but showed that water was excluded from the hydrophobic interface between the protein and the DNA bases (Clore *et al.*, 1994). Although many of the existing crystal structures of protein–DNA complexes support the general view that water molecules are often integral components of the specific recognition between the protein and the target DNA, this solution structure provides an important example of exclusion of water molecules from the specificity determinants. In the GATA-1–DNA complex, however, water molecules do mediate non-specific binding of the protein to the DNA backbone. It appears, not surprisingly, that water molecules play a variety of roles in the mediation of protein–DNA interactions and that these roles are specific to each particular case.

23.5.6.3. Cooperativity in dimeric haemoglobin

The X-ray crystal structures of liganded and unliganded dimeric haemoglobin from *Scapharca inaequivalvis* have revealed that water molecules at the dimer interface form an

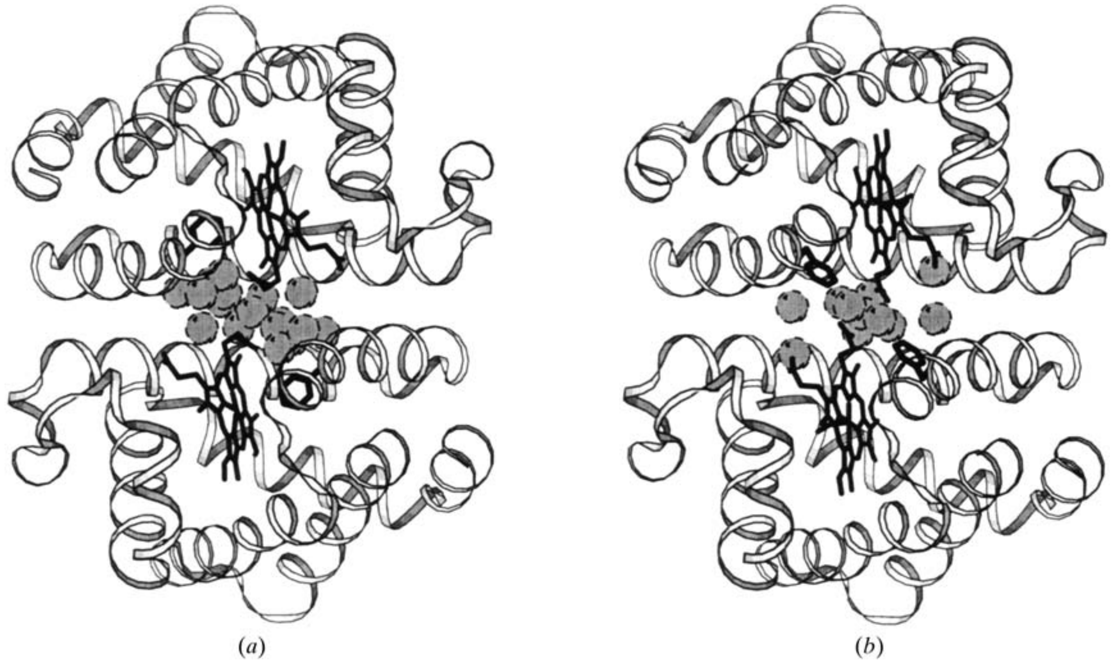


Figure 23.5.6.1
Scapharca HbI interface water molecules. (*a*) Deoxy-HbI at 1.6 Å resolution (PDB code 3SDH) and (*b*) HbI-CO at 1.4 Å resolution (PDB code 4SDH). Included is a ribbon diagram showing the tertiary structure of each subunit, bond representations for the haem group and Phe97 side chain, and spheres representing the approximate van der Waals radii of oxygen atoms for core interface water molecules. Note the cluster of 17 ordered water molecules in the interface of deoxy-HbI for which Phe97 is packed in the haem pocket. Upon ligation, by either CO or O_2, Phe97 is extruded into the interface and disrupts this water cluster, expelling six water molecules from the interface. These plots were produced with the program *MOLSCRIPT* (Kraulis, 1991). Reprinted with permission from Royer *et al.* (1997). Copyright (1997) The American Society for Biochemistry & Molecular Biology.

integral part of the cooperativity mechanism in this system (Condon & Royer, 1994; Royer, 1994). The binding of oxygen to one of the monomers causes little rearrangement of quaternary structure. It does, instead, displace the side chain of Phe97 which, in the low-affinity deoxy form, packs in the haem pocket (Royer *et al.*, 1990). Phe97 in the deoxy form lowers the oxygen affinity by restricting movement of the iron atom into the haem plane (Royer, 1994). Upon oxygen binding, Phe97 flips to the dimer interface, removing six out of the 17 water molecules that are found in the deoxy form (Fig. 23.5.6.1). The resultant destabilization of the water clusters found between the two subunits facilitates the flipping of Phe97 in the other subunit, with a concomitant increase in oxygen affinity of the haem in the second subunit (Pardanani *et al.*, 1997; Royer *et al.*, 1997).

In each of the monomeric subunits, Thr72 is positioned to form a hydrogen bond with a water molecule at the periphery of the deoxy dimer interface (not shown in Fig. 23.5.6.1). In effect, this interaction caps the water cluster on either side of the interface, presumably helping to stabilize these well ordered water molecules. The isosteric mutation Thr72 to Val was designed to test the importance of this interaction to the stability of the water cluster in the low-affinity haemoglobin dimer and the resultant effect on ligand affinity and cooperativity (Royer *et al.*, 1996). The crystal structure of the T72V mutant was solved to 1.6 Å resolution. This crystal structure reveals that the only significant difference between the mutant and wild-type proteins is the loss of the two water molecules that directly hydrogen-bond to Thr72 in each of the wild-type subunits. Furthermore, there is a significant increase in both activity and cooperativity resulting from the mutation (Royer *et al.*, 1996). The authors conclude that, as a result of the mutation, the loss of two water molecules in the interface cluster is sufficient to alter the balance between the low- and high-affinity forms of the protein. This result demonstrates that water molecules are key mediators of information transfer between the haems in the two subunits in dimeric haemoglobin and that their precise positioning and interactions with protein atoms are crucial in maintaining the chemical balance required for biological function.

23.5.6.4. *Summary*

The few examples illustrated above provide diverse views of the ways in which Nature can use water molecules as integral parts of macromolecular interactions. Water molecules can be involved in specificity and recognition, in thermodynamics of binding and affinity, in the cooperative behaviour of allosteric proteins, and in catalysis. Not only do the specific examples illustrate general roles possible for water molecules in the context of a given type of macromolecule, such as proteins or nucleic acids, but they are often representative of any macromolecular system. For example, the role of water in recognition and specificity illustrated above for protein–DNA interactions has also been observed in the L-arabinose-binding protein interaction with specific sugar molecules (Quiocho *et al.*, 1989). Clearly, water molecules are involved so intimately, and in so many different ways, with the formation of molecular complexes that it is not possible to understand the formation process and the function of the complex without taking into account the role of this universal solvent.

23.5.7. Conclusions and future perspectives

The aim of this chapter was to provide a general overview of the available crystallographic information on the roles of water molecules in their interactions with macromolecules. To achieve this aim, the focus has been on representative examples rather than an exhaustive review of the literature. The classification of water molecules according to location and frequency of occurrence among related proteins, or independently solved crystal structures of a given protein, is a crucial element in determining their functional roles.

It has become clear that water molecules involved in crystal contacts can occur virtually anywhere on the protein surface, as exemplified in the case of T4 lysozyme, and that they have properties similar to the majority of surface water molecules within the context of the crystal. Therefore, it is possible to conclude that for the majority of cases, the crystal contacts between proteins involved in the various studies discussed in this review do not significantly influence the general conclusions drawn.

Water molecules bind to proteins so as to satisfy the hydrogen-bonding potential of protein atoms that are not part of the intramolecular hydrogen-bonding pattern within the native structure. At the primary level, the hydrogen bonding is such as to follow the stereochemical requirements of the individual atom in question, in a manner similar to that occurring for the same atom in small molecules. At the secondary-structure level, these positions tend to provide extensions of α-helices or β-sheets as well as to solvate protein atoms in exposed turns. At the tertiary level, they occur more favourably in grooves or cavities within the protein.

Internal or buried water molecules are found to bridge between domains of a single monomer or bring together different secondary-structure elements within a given domain. They have also been observed in the binding sites of proteins where they fine-tune the shape or electrostatic complementarity towards the substrate or ligand. In general, buried water molecules occur in cavities within the protein, making multiple hydrogen bonds with protein atoms that are likely to be conserved among members of a given family. These water molecules are themselves conserved and must be considered an integral part of the protein architecture. They are often connected to bulk water through water channels leading to the protein surface.

Surface water molecules play important roles in protein dynamics, catalysis, thermodynamics of binding, and in mediation of cooperativity, metal binding, recognition and specificity. Representative examples of water molecules in each of these different roles are discussed in the present review. Some of the surface water molecules are found to be conserved within families of proteins, particularly when they are involved in one of the specific roles mentioned above.

In addition to the commonly observed features in crystal structures of proteins solved to around 2 Å resolution, the crystal structures of crambin and BPTI, both solved to 1 Å resolution, provide examples of the type of information only available at very high resolution. This includes the arrangement of water molecules into pentagonal rings around hydrophobic side chains and the occupancy of water molecules on protein surfaces.

A cohesive picture has emerged of the locations of well ordered water molecules on protein surfaces and their functional roles. Currently, there is a good structural view of the protein atoms as well as of the structure of water molecules associated with the protein. The question now is how this information can be used in predicting *a priori* where water molecules will be involved in important structural or functional roles. While having information on the ordered water molecules associated with protein surfaces represents significant progress toward the ultimate goal of understanding the global thermodynamic and kinetic picture of molecular processes in water, the entire system is still not understood. For instance, how does the bulk water influence the dynamic and thermodynamic processes in which the protein and ordered water molecules are involved? Furthermore, what is the importance of solutes normally found in the biological environment where proteins and other macromolecules exert their function? Knowledge must continue to expand toward an understanding of the complete system. Meanwhile, the present models can go a long way toward successful practical applications in protein engineering and ligand design. In order to improve these models, the information accumulated so far can be combined with empirical results and theoretical models to expand the understanding of the first principles underlying biological processes. Building the bridge between empirical observation and first principles is an iterative process still in its infancy.

We wish to thank Martin Karplus and Gregory A. Petsko for years of support and discussions that ultimately contributed to the integrity of this review. During the writing of this review Carla Mattos was supported by the American Cancer Society Postdoctoral Fellowship grant No. PF-4331.

References

Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A. & Ringe, D. (1996). *An experimental approach to mapping the binding surfaces of crystalline proteins. J. Phys. Chem.* **100**, 2605–2611.

Badger, J. (1993). *Multiple hydration layers in cubic insulin crystals. Biophys. J.* **65**, 1656–1659.

Baker, E. N. & Hubbard, R. E. (1984). *Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol.* **44**, 97–179.

Beglov, D. & Roux, B. (1997). *An integral equation to describe the solvation of polar molecules in liquid water. J. Phys. Chem.* **101**, 7821–7826.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bhat, T. N., Bentley, G. A., Boulot, G., Greene, M. I., Tello, D., Dall'Acqua, W., Souchon, H., Schwarz, F. P., Maiuzza, R. A. & Poljak, R. J. (1994). *Bound water molecules and conformational stabilization help mediate an antigen–antibody association. Proc. Natl Acad. Sci. USA*, **91**, 1089–1093.

Blake, C. C. F., Pulford, W. C. A. & Artymiuk, P. J. (1983). *X-ray studies of water in crystals of lysozyme. J. Mol. Biol.* **167**, 693–723.

Brooks, C. L. & Karplus, M. (1989). *Solvent effects on protein motion and protein effects on solvent motion. Dynamics of the active site region of lysozyme. J. Mol. Biol.* **208**, 159–181.

Bryant, R. G. (1996). *The dynamics of water–protein interactions. Annu. Rev. Biophys. Biomol. Struct.* **25**, 29–53.

Chervenak, M. C. & Toone, E. J. (1994). *A direct measure of the contribution of solvent reorganization to the enthalpy of ligand binding. J. Am. Chem. Soc.* **116**, 10533–10539.

Clackson, T. & Wells, J. T. (1995). *A hot spot of binding energy in a hormone–receptor interface. Science*, **267**, 383–386.

Clark, K. L., Halay, E. D., Lai, E. & Burley, S. K. (1993). *Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature (London)*, **364**, 412–420.

Clore, G. M., Bax, A., Omichinski, J. G. & Gronenborn, A. M. (1994). *Localization of bound water in the solution structure of a complex of the erythroid transcription factor GATA-1 with DNA. Curr. Biol.* **2**, 89–94.

Condon, P. & Royer, W. (1994). *Crystal structure of oxygenated Scapharca dimeric hemoglobin at 1.7 Å resolution. J. Biol. Chem.* **269**, 25259–25267.

Dechene, M., Wink, G., Smith, M., Swartz, P. & Mattos, C. (2009). *Multiple solvent crystal structures of ribonuclease A: an assessment of the method. Proteins Struct. Funct. Bioinform.* **76**, 861–881.

Deisenhofer, J. & Steigemann, W. (1975). *Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1.5 Å resolution. Acta Cryst.* B**31**, 238–250.

Edsall, J. T. & McKenzie, H. A. (1978). *Water and proteins. I. The significance and structure of water; its interaction with electrolytes and non-electrolytes. Adv. Biophys.* **10**, 137–207.

Edsall, J. T. & McKenzie, H. A. (1983). *Water and proteins. II. The location and dynamics of water in protein systems and its relation to their stability and properties. Adv. Biophys.* **16**, 53–183.

Gunsteren, W. F. van, Luque, F. J., Timms, D. & Torda, A. E. (1994). *Molecular mechanics in biology: from structure to function, taking account of solvation. Annu. Rev. Biophys. Biomol. Struct.* **23**, 847–863.

Hayward, S., Kitao, A., Hirata, F. & Go, N. (1993). *Effect of solvent on collective motions in globular protein. J. Mol. Biol.* **234**, 1207–1217.

Hendrickson, W. A. & Teeter, M. M. (1981). *Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur. Nature (London)*, **290**, 107–113.

Hendsch, Z. S., Jonsson, T., Sauer, R. T. & Tidor, B. (1996). *Protein stabilization by removal of unsatisfied polar groups: computational approaches and experimental tests. Biochemistry*, **35**, 7621–7625.

Hendsch, Z. S. & Tidor, B. (1994). *Do salt bridges stabilize proteins? A continuum electrostatic analysis. Protein Sci.* **3**, 211–226.

Herron, J. N., Terry, A. H., Johnston, S., He, S.-M., Guddat, L. W., Voss, E. W. & Edmundson, A. B. (1994). *High resolution structures of the 4-4-20 Fab–fluorescein complex in two solvent systems: effects of solvent on structure and antigen-binding affinity. Biophys. J.* **67**, 2167–2175.

Holdgate, G., Tunnicliffe, A., Ward, W. H. J., Weston, S. A., Rosenbrock, G., Barth, P. T., Taylor, I. W. F., Pauptit, R. A. & Timms, D. (1997). *The entropic penalty of ordered water accounts for weaker binding of the antibiotic Novobiocin to a resistant mutant of DNA gyrase: a thermodynamic and crystallographic study. Biochemistry*, **36**, 9663–9673.

Hubbard, S. J., Gross, K.-H. & Argos, P. (1994). *Intramolecular cavities in globular proteins. Protein Eng.* **7**, 613–626.

Jiang, J.-S. & Brünger, A. (1994). *Protein hydration observed by X-ray diffraction. Solvation properties of penicillopepsin and neuraminidase crystal structures. J. Mol. Biol.* **243**, 100–115.

Karplus, P. A. & Faerman, C. (1994). *Ordered water in macromolecular structure. Curr. Opin. Struct. Biol.* **4**, 770–776.

Kauzmann, W. (1959). *Some factors in the interpretation of protein denaturation. Adv. Protein Chem.* **14**, 1–63.

Kendrew, J. C. (1963). *Myoglobin and the structure of proteins. Science*, **139**, 1259–1266.

Komives, E. A., Lougheed, J. C., Liu, K., Sugio, S., Zhang, Z., Petsko, G. A. & Ringe, D. (1995). *The structural basis for pseudoreversion of the E165D lesion by the secondary S96P mutation in triosephosphate isomerase depends on the positions of active site water molecules. Biochemistry*, **34**, 13612–13621.

Kossiakoff, A. A., Sintchak, M. D., Shpungin, J. & Presta, L. G. (1992). *Analysis of solvent structure in proteins using neutron $D_2O - H_2O$ solvent maps: pattern of primary and secondary hydration of trypsin. Proteins Struct. Funct. Genet.* **12**, 223–236.

Kraulis, P. J. (1991). *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst.* **24**, 946–950.

Kuhn, L., Siani, M. A., Pique, M. E., Fisher, C. L., Getzoff, E. D. & Tainer, J. A. (1992). *The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. J. Mol. Biol.* **228**, 13–22.

Kuhn, L. A., Swanson, C. A., Pique, M. E., Tainer, J. A. & Getzoff, E. D. (1995). *Atomic and residue hydrophilicity in the context of folded protein structures. Proteins Struct. Funct. Genet.* **23**, 536–547.

Ladbury, J. E. (1996). *Just add water! The effect of water on the specificity of protein–ligand binding sites and its potential application to drug design. Chem. Biol.* **3**, 973–980.

Lam, P. Y. S., Jadhav, P. K., Eyermann, C. J., Hodge, C. N., Ru, Y., Bacheler, L. T., Meek, J. L., Otto, M. J., Rayner, M. M., Wong, Y. N., Chang, C.-H., Weber, P. C., Jackson, D. A., Sharpe, T. R. & Erickson-Vitanen, S. (1994). *Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science*, **263**, 380–384.

Lazaridis, T., Archontis, G. & Karplus, M. (1995). *Enthalpic contribution to protein stability: insights from atom-based calculations and statistical mechanics. Adv. Protein Chem.* **47**, 231–306.

Levitt, M. & Park, B. H. (1993). *Water: now you see it, now you don't. Structure*, **1**, 223–226.

Loris, R., Langhorst, U., De Vos, S., Decanniere, K., Bouckaert, J., Maes, D., Transue, T. R. & Steyaert, J. (1999). *Conserved water molecules in a large family of microbial ribonucleases. Proteins Struct. Funct. Genet.* **36**, 117–134.

Loris, R., Stas, P. P. G. & Wyns, L. (1994). *Conserved waters in legume lectin crystal structures. The importance of bound water for the sequence–structure relationship within the legume lectin family. J. Biol. Chem.* **269**, 26722–26733.

Lounnas, V. & Pettitt, B. M. (1994). *Distribution function implied dynamics versus residence times and correlations: solvation shells of myoglobin. Proteins Struct. Funct. Genet.* **18**, 148–160.

Lounnas, V., Pettitt, B. M. & Phillips, G. N. Jr (1994). *A global model of the protein–solvent interface. Biophys. J.* **66**, 601–614.

McDowell, R. S. & Kossiakoff, A. A. (1995). *A comparison of neutron diffraction and molecular dynamics structures: hydroxyl group and water molecule orientations in trypsin. J. Mol. Biol.* **250**, 553–570.

Malin, R., Zielenkiewicz, P. & Saenger, W. (1991). *Structurally conserved water molecules in ribonuclease T1. J. Biol. Chem.* **266**, 4848–4852.

Mattos, C. (2002). *Protein–water interactions in a dynamic world. Trends Biochem. Sci.* **27**, 203–208.

Mattos, C., Bellamacina, C., Peisach, E., Pereira, A., Vitkup, D., Petsko, G. A. & Ringe, D. (2006). *Multiple solvent crystal structures: probing binding sites, plasticity and hydration. J. Mol. Biol.* **357**, 1471–1482.

Mattos, C. & Clark, A. C. (2008). *Minimizing frustration by folding in an aqueous environment. Arch. Biochem. Biophys.* **469**, 118–131.

Mattos, C., Giammona, D. A., Petsko, G. A. & Ringe, D. (1995). *Structural analysis of the active site of porcine pancreatic elastase based on the X-ray crystal structures of complexes with trifluoroacetyl-dipeptide-anilide inhibitors. Biochemistry*, **34**, 3193–3203.

Mattos, C., Rasmussen, B., Ding, X., Petsko, G. A. & Ringe, D. (1994). *Analogous inhibitors of elastase do not always bind analogously. Nat. Struct. Biol.* **1**, 55–58.

Mattos, C. & Ringe, D. (1996). *Locating and characterizing binding sites on proteins. Nat. Biotech.* **14**, 595–599.

Meiering, E. M. & Wagner, G. (1995). *Detection of long-lived bound water molecules in complexes of human dihydrofolate reductase with methotrexate and NADPH. J. Mol. Biol.* **247**, 294–308.

Meyer, E. (1992). *Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. Protein Sci.* **1**, 1543–1562.

Meyer, E., Cole, G., Radhakrishnan, R. & Epp, O. (1988). *Structure of native porcine pancreatic elastase at 1.65 Å resolution. Acta Cryst.* B**44**, 26–38.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). *Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. J. Phys. Chem.* **79**, 2361–2381.

Morton, C. J. & Ladbury, J. E. (1996). *Water mediated protein–DNA interactions: the relationship of thermodynamics to structural detail. Protein Sci.* **5**, 2115–2118.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). *Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. Proteins Struct. Funct. Genet.* **11**, 281–296.

Omichinski, J. G., Clore, G. M., Schaad, O., Felsenfeld, G., Trainor, C., Appella, E., Stahl, S. J. & Gronenborn, A. (1993). *NMR structure of a specific DNA complex of Zn-containing DNA binding domain of GATA-1. Science*, **261**, 438–446.

Oprea, T. I., Hummer, G. & Garcia, A. E. (1997). *Identification of a functional water channel in cytochrome P450 enzymes. Proc. Natl Acad. Sci. USA*, **94**, 2133–2138.

Otting, G., Liepinsh, E. & Wuthrich, K. (1991). *Protein hydration in aqueous solution. Science*, **254**, 974–980.

Otting, G. & Wuthrich, K. (1989). *Studies of protein hydration in aqueous solution by direct NMR observation of individual protein-bound water molecules. J. Am. Chem. Soc.* **111**, 1871–1875.

Otwinowski, Z., Schevitz, R. W., Zhang, R.-G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., Luisi, B. F. & Sigler, P. B. (1988). *Crystal structure of trp repressor/operator complex at atomic resolution. Nature (London)*, **335**, 321–329.

Pardanani, A., Gibson, Q. H., Colotti, G. & Royer, W. E. (1997). *Mutation of residue Phe97 to Leu disrupts the central allosteric pathway in Scapharca dimeric hemoglobin. J. Biol. Chem.* **272**, 13171–13179.

Pletinckx, J., Steyaert, J., Zegers, I., Choe, H.-W., Heinemann, U. & Wyns, L. (1994). *Crystallographic study of Glu58Ala RNase T1-2′-guanosine monophosphate at 1.9 Å resolution. Biochemistry*, **33**, 1654–1662.

Pomes, R. & Roux, B. (1996). *Structure and dynamics of a proton wire: a theoretical study of H^+ translocation along the single-file water chain in the gramicidin A channel. Biophys. J.* **71**, 19–39.

Poormina, C. S. & Dean, P. M. (1995*a*). *Hydration in drug design. 3. Conserved water molecules at the ligand-binding sites of homologous proteins. J. Comput.-Aided Mol. Des.* **9**, 521–531.

Poormina, C. S. & Dean, P. M. (1995*b*). *Hydration in drug design. 1. Multiple hydrogen-bonding features of water molecules in mediating protein-ligand interactions. J. Comput.-Aided Mol. Des.* **9**, 500–512.

Poormina, C. S. & Dean, P. M. (1995*c*). *Hydration in drug design. 2. Influence of local site surface shape on water binding. J. Comput.-Aided Mol. Des.* **9**, 513–520.

Privé, G. G., Milburn, M. V., Tong, L., DeVos, A. M., Yamaizumi, Z., Nishimura, S. & Kim, S. H. (1992). *X-ray crystal structures of transforming p21 ras mutants suggest a transition-state stabilization mechanism for GTP hydrolysis. Proc. Natl Acad. Sci. USA*, **89**, 3649–3653.

Quiocho, F. A., Wilson, D. K. & Vyas, N. K. (1989). *Substrate specificity and affinity of a protein modulated by bound water molecules. Nature (London)*, **340**, 404–407.

Rand, R. P. (1992). *Raising water to new heights. Science*, **256**, 618.

Rashin, A. A., Iofin, M. & Honig, B. (1986). *Internal cavities and buried waters in globular proteins. Biochemistry*, **25**, 3619–3625.

Ringe, D. (1995). *What makes a binding site a binding site? Curr. Opin. Struct. Biol.* **5**, 825–829.

Robinson, C. R. & Siglar, S. G. (1993). *Molecular recognition mediated by bound water. A mechanism for star activity of the restriction endonuclease EcoRI. J. Mol. Biol.* **234**, 302–306.

Roe, S. M. & Teeter, M. M. (1993). *Patterns for prediction of hydration around polar residues in proteins. J. Mol. Biol.* **229**, 419–427.

Roux, B., Nina, M., Pomes, R. & Smith, J. C. (1996). *Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy pertubation study. Biophys. J.* **72**, 670–681.

Royer, W. (1994). *High-resolution crystallographic analysis of a co-operative dimeric hemoglobin. J. Mol. Biol.* **235**, 657–681.

Royer, W. E., Fox, R. A. & Smith, F. R. (1997). *Ligand linked assembly of Scapharca dimeric hemoglobin. J. Biol. Chem.* **272**, 5689–5694.

Royer, W. E. Jr, Hendrickson, W. A. & Chiancone, E. (1990). *Structural transitions upon ligand binding in a cooperative dimeric hemoglobin. Science*, **249**, 518–521.

Royer, W. E., Pardanani, A., Gibson, Q. H., Peterson, E. S. & Friedman, J. M. (1996). *Ordered water molecules as key allosteric mediators in a cooperative dimeric hemoglobin. Proc. Natl Acad. Sci. USA*, **93**, 14526–14531.

Savage, H. (1986). *Water structure in vitamin B_{12} coenzyme crystals. Biophys. J.* **50**, 967–980.

Savage, H. & Wlodawer, A. (1986). *Determination of water structure around biomolecules using X-ray and neutron diffraction methods. Methods Enzymol.* **127**, 162–183.

Shakked, Z., Guzikevich-Guerstein, G., Frolow, F., Rabinovich, D., Joachimiak, A. & Sigler, P. B. (1994). *Determinants of repressor/operator recognition from the structure of the trp operator binding site. Nature (London)*, **368**, 469–473.

Shaltiel, S., Cox, S. & Taylor, S. (1998). *Conserved water molecules contribute to the extensive network of interactions at the active site of protein kinase A. Proc. Natl Acad. Sci. USA*, **95**, 484–491.

Shpungin, J. & Kossiakoff, A. A. (1986). *A method of solvent structure analysis for proteins using $D_2O - H_2O$ neutron difference maps. Methods Enzymol.* **127**, 329–342.

Singer, P., Smalas, A., Carty, R. P., Mangel, W. F. & Sweet, R. M. (1993). *The hydrolytic water molecule in trypsin, revealed by time-resolved Laue crystallography. Science*, **259**, 669–673.

Sreenivasan, U. & Axelsen, P. H. (1992). *Buried water in homologous serine proteases. Biochemistry*, **31**, 12785–12791.

Teeter, M. M. (1984). *Water structure of a hydrophobic protein at atomic resolution: pentagon rings of water molecules in crystals of crambin. Proc. Natl Acad. Sci. USA*, **81**, 6014–6018.

Teeter, M. M. (1991). *Water–protein interactions: theory and experiment. Annu. Rev. Biophys. Biophys. Chem.* **20**, 577–600.

Teeter, M. M. & Hendrickson, W. A. (1979). *Highly ordered crystals of the plant seed protein crambin. J. Mol. Biol.* **127**, 219–223.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1988). *Distribution of water around amino acid residues in proteins. J. Mol. Biol.* **202**, 637–657.

Thanki, N., Thornton, J. M. & Goodfellow, J. M. (1990). *Influence of secondary structure on the hydration of serine, threonine and tyrosine residues in proteins. Protein Eng.* **3**, 495–508.

Thanki, N., Umrania, Y., Thornton, J. M. & Goodfellow, J. M. (1991). *Analysis of protein main-chain solvation as a function of secondary structure. J. Mol. Biol.* **221**, 669–691.

Walshaw, J. & Goodfellow, J. M. (1993). *Distribution of solvent molecules around apolar side-chains in protein crystals. J. Mol. Biol.* **231**, 392–414.

Weaver, L. & Matthews, B. (1987). *Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. J. Mol. Biol.* **193**, 189–199.

Williams, M. A., Goodfellow, J. M. & Thornton, J. M. (1994). *Buried waters and internal cavities in monomeric proteins. Protein Sci.* **3**, 1224–1235.

Wlodawer, A., Walter, J., Huber, R. & Sjolin, L. (1984). *Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and X-ray refinement of crystal form II. J. Mol. Biol.* **180**, 301–329.

Zegers, I., Maes, D., Dao-Thi, M.-H., Poortmans, F., Palmer, R. & Wyns, L. (1994). *The structures of RNase A complexed with 3′-CMP and d(CpA): active site conformation and conserved water molecules. Protein Sci.* **3**, 2322–2339.

Zhang, X.-J. & Matthews, B. W. (1994). *Conservation of solvent-binding sites in 10 crystal forms of T4 lysozyme. Protein Sci.* **3**, 1031–1039.

International Tables for Crystallography (2012). Vol. F, Chapter 23.6, pp. 821–826.

Chapter 23.6. Halogen interactions in biomolecular crystal structures

M. J. Vallejos, P. Auffinger and P. S. Ho

23.6.1. Introduction

Halogens are uncommon substituents in biological molecules. It is not surprising, therefore, that molecular interactions involving covalent halogen compounds are not foremost in the minds of most biologists. There is, however, an increasing appreciation that halogenated compounds play important roles in biology, starting with the naturally halogenated hormone thyroxine (Fig. 23.6.1.1*a*) (Wojtczak *et al.*, 2001), and extending to bioactive secondary metabolites isolated as natural products (Gribble, 2004) and a growing list of synthetic compounds that serve as inhibitors against a broad range of enzymes. Proteins and nucleic acids are naturally halogenated through oxidative processes associated with inflammatory response mechanisms (van Pée & Unversucht, 2003; Shen *et al.*, 2001), with chlorinated and brominated amino acids in proteins correlated with various respiratory ailments and allergic responses (Buss *et al.*, 2003; Wu *et al.*, 2000). In structural biology, it is now common to add electron-rich halogens (particularly Br and I) to help solve the phase problem in crystallography, or magnetically active nuclei (^{19}F) to facilitate structure determination by NMR. In nucleic acids, halogenated nucleotide bases can be incorporated in synthetic oligonucleotides to help determine the structures of DNA and RNA, as well as of protein–nucleic acid complexes. Halogenated amino acids such as bromophenylalanine can now be introduced site-specifically into proteins through *in vivo* incorporation (Wang *et al.*, 2006), which can also facilitate phasing of X-ray diffraction data for protein crystals (Xie *et al.*, 2004), while synthetic fluorinated amino acids have been used to facilitate structural studies by ^{19}F NMR (Hammill *et al.*, 2007). Together, these various systems have contributed to the recent growth in the number of halogens and halogen interactions observed in biomolecular structures.

For this particular discussion, we will focus on the molecular interactions of halogenated compounds seen in structures determined by single-crystal X-ray diffraction rather than by NMR, primarily because of the strong dependence on the distance and angle relationships required to categorize and understand such interactions (NMR structures are highly dependent on appropriate force fields to define conformation; unfortunately, current force fields do not properly treat interactions involving halogens). A survey of the 16 June 2009 release of the Protein Data Bank (PDB; Berman *et al.*, 2000) showed that there were 2670 entries containing a halogen atom (Table 23.6.1.1). In other words, nearly 5% of the more than 58 000 entries from this release contained at least one covalent halogen, indicating that halogens are an important component in biomolecular structures.

23.6.2. Classical treatment of halogen interactions

23.6.2.1. Non-specific hydrophobic effects

Halogen interactions in proteins and nucleic acid structures have classically been treated non-electrostatically. Indeed, most quantitative structure–activity relationships (QSARs) consider only the steric and lipophilic contributions of halogens when trying to predict their effect on ligand binding. The majority of interactions can be classified in this sense; non-electrostatic interactions would predict distances between halogen and other non-hydrogen atoms to be greater than or equal to the sum of their respective van der Waals radii ($\sum r_{\rm vdW}$), with the halogen atoms sitting in hydrophobic pockets. An example of this is seen in the binding of dibromomethane to methane monooxygenase hydrolase (Egner *et al.*, 2005), where the Br atoms of the bound ligand are surrounded primarily by the hydrophobic side chains (Leu, Val and Phe) of the protein (Whittington *et al.*, 2001). The non-electrostatic components of molecular halogens can be characterized by the size (as measured by $r_{\rm vdW}$) and hydrophobicity (as measured by the difference in partition coefficients and the associated molecular solvation free energy for the accessible surface) of the respective halogen atoms (Table 23.6.2.1).

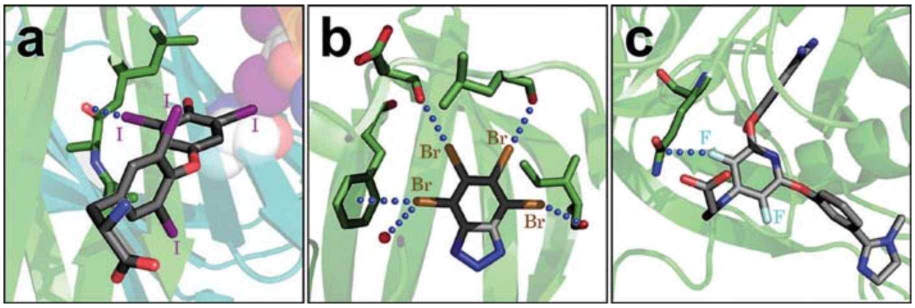


Figure 23.6.1.1
Examples of electrostatic interactions of halogenated ligands with proteins. The C atoms of the protein residues are shown in green and those of the bound ligands in black. (*a*) An I atom from tetraiodothyroxine (violet) is shown with its close (3.07 Å) *X*-bonding interaction (blue dots) to the carbonyl O atom of Ala109 of transthyretin (Wojtczak *et al.*, 2001). (*b*) The Br atoms (orange) of tetrabromobenzotriazole form short contacts to the carbonyl O atoms of Glu81 (3.0 Å), Leu83 (2.9 Å) and Ile10 (3.3 Å), and to the aromatic ring of Phe80 (3.5 Å to the centre), of phospho-CDK2/cyclin A (De Moliner *et al.*, 2003). There is an additional interaction with a water molecule (2.91 Å). (*c*) An F atom (cyan) from the inhibitor ZK-807834 interacts with the amide C atom of the Gln192 side chain in the ligand-binding site of the clotting factor Xa (Adler *et al.*, 2000).

23.6.2.2. Substituent effects

The classical approach to the electrostatic effects of halogens is to consider them as electron-withdrawing substituents: I, Br, Cl and F atoms are strong electron-withdrawing groups, relative to hydrogen, in their effects on the reactivities of *para*- and *meta*-substituted benzoic acids [reflected in the so-called Hammett σ constants (Hammett, 1937), Table 23.6.2.1]. However, such substituent effects do not tell us anything about how halogens themselves participate in directing non-covalent interactions. Our basic understanding is that halogens are electron-rich and therefore we would expect them to act as Lewis bases, serving, for example, as hydrogen-bond acceptors (like carbonyl O or

Copyright © 2012 International Union of Crystallography

Table 23.6.1.1
Interactions in crystal structures in the Protein Data Bank

The numbers of interactions (N_I) for molecular halogens with oxygen (O), nitrogen (N), sulfur (S) and carbon atoms (C) are tabulated for single-crystal structures at resolutions of 3.0 Å or better from the 16 June 2009 release of the PDB (containing 58 236 total entries). A contact is considered to be an interaction if the distance between the halogen and its partner is between 2.2 Å and the van der Waals contact distance ($D_{\max}$†). O-atom partners are counted separately for non-water and water interactions, while interactions with non-aromatic atoms are counted separately from those involving aromatic atoms.

	F		Cl		Br		I	
Acceptor atom	$D_{\max}$ (Å)	N_I	$D_{\max}$ (Å)	N_I	$D_{\max}$ (Å)	N_I	$D_{\max}$ (Å)	N_I
Non-aromatic (backbones of polypeptides and polynucleotides, or side chains of non-aromatic amino acids)								
O (non-water)	2.99	1219	3.27	241	3.37	198	3.50	285
O (water)	3.52	1454	3.80	263	3.90	1110	4.03	637
N	3.02	977	3.30	206	3.40	78	3.53	139
S	3.27	45	3.55	39	3.65	13	3.78	16
C	3.17	1466	3.45	511	3.55	338	3.68	472
Aromatic (side chains of Phe, Tyr, Trp and His in proteins, and the bases of nucleotides)								
O	2.99	42	3.27	22	3.37	58	3.50	8
N	3.02	11	3.30	9	3.40	16	3.53	0
C	3.17	174	3.45	242	3.55	237	3.68	77
Total	5388		1533		2048		1634	

† $D_{\max}$ is the sum of the van der Waals radii [$r_{\rm vdW}$; Bondi (1964)] for the halogen and its interaction partner. In the case of water, this is the sum of the length of the O—H bond (0.98 Å) plus $r_{\rm vdW}$ of the H atom (1.09 Å) and $r_{\rm vdW}$ of the respective halogen. The longer $D_{\max}$ for interactions with water allows the halogens to serve as hydrogen-bond acceptors.

Table 23.6.2.1
Physical properties of molecular halogens

Comparative values are shown for each halogen for their van der Waals radii† ($r_{\rm vdW}$), their contribution to partition coefficients in water *versus* octanol of the respective halobenzene *versus* unsubstituted benzene ($P_{x\varphi}$ and P_{φ}, respectively‡), the associated contributions to the solvation free energies§ (SFEs), and the substituent effects for *para*- or *meta*-substituted benzoic acids, as reflected in the Hammett constants¶ (σ_p and σ_m, respectively).

Halogen (X)	$r_{\rm vdW}$ (Å)	$\log(P_{x\varphi}) - \log(P_{\varphi})$	SFE (cal Å^{-2})	σ_p	σ_m
F	1.47	0.12	11.9	0.062	0.337
Cl	1.75	0.34	13.2	0.227	0.373
Br	1.83	0.84	37.6	0.232	0.393
I	1.98	1.10	41.5	0.276	0.353

† From Bondi (1964). ‡ Defined as the base-10 logarithm of the partition (P) from water to octanol of the halogenated benzene ($X\varphi$, where X is the halogen) minus that of benzene (φ). Partition coefficients from Hansch (1979). § Calculated as $-2.303RT[\log(P_{x\varphi}) - \log(P_{\varphi})]$ divided by the solvent-accessible surface of the halogen in the respective halobenzene model. ¶ From Hammett (1937). Positive values reflect the electron-withdrawing capacity of the substituent.

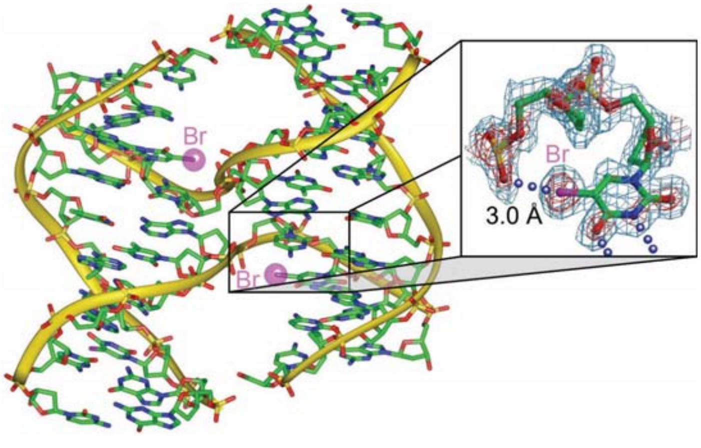


Figure 23.6.3.1
The single-crystal structure of the DNA sequence d(CCAGTAC-brUGG), where brU is 5-bromouracil (Br atom shown as a sphere), as a four-stranded Holliday junction. The inset shows the electron-density map ($2F_o - F_c$ map, contoured at 1.5σ in blue and 3σ in red) around the brU residue, with the short contact between the Br atom and the phosphate O atom at the tight U-turn of the DNA junction backbone indicated.

imine N atoms). Indeed, F atoms can typically replace hydroxyl groups as hydrogen-bond acceptors in carbohydrates (Scott & Viola, 1998; Lougheed *et al.*, 1999; Ly *et al.*, 2000; Hart *et al.*, 2000).

23.6.3. Electrostatic molecular halogen interactions

Non-covalent halogen interactions are recognized as being highly diverse, extending well beyond the classical properties of size, lipophilicity and substituent effects. The remainder of this review will focus on a class of electrostatic interactions called halogen bonds (X bonds) and other similar electrostatic interactions, primarily because these are specific interactions that have been shown to affect molecular conformations and ligand binding in single-crystal structures of proteins and nucleic acids.

23.6.3.1. Biological halogen bonds

The growing use of halogens for crystallographic phasing (halide soaks notwithstanding) and in natural product and drug development means that structural biologists must become more aware of the non-covalent molecular interactions involving the Group VII elements. In many cases, halogens are introduced as ‘neutral’ substitutions, but they are not altogether unobtrusive: there are now several examples where a halogenated nucleotide base has been seen to very dramatically affect the overall structure of DNA. In one such example, a brominated uracil base induced a decanucleotide sequence to crystallize as a four-stranded Holliday junction (Fig. 23.6.3.1), while the analogous unbrominated and methylated sequences crystallized as standard B-DNA double helices (Hays *et al.*, 2003). Although we did not recognize this as an X bond at the time, the Br atom was seen to form a stabilizing interaction with the negatively charged phosphate O atom to force the DNA backbone into a tight U-turn and, as a result, to help form the four-stranded complex. In fact, we and others (Muzet *et al.*, 2003) had thought that such short halogen–oxygen distances represented a completely new type of interaction, until an internet search led to the lecture by Odd Hassel for the 1969 Nobel Prize in Chemistry (Hassel, 1972). In short, there is ‘nothing new under the sun’ [Ecclesiastes 1:9–14].

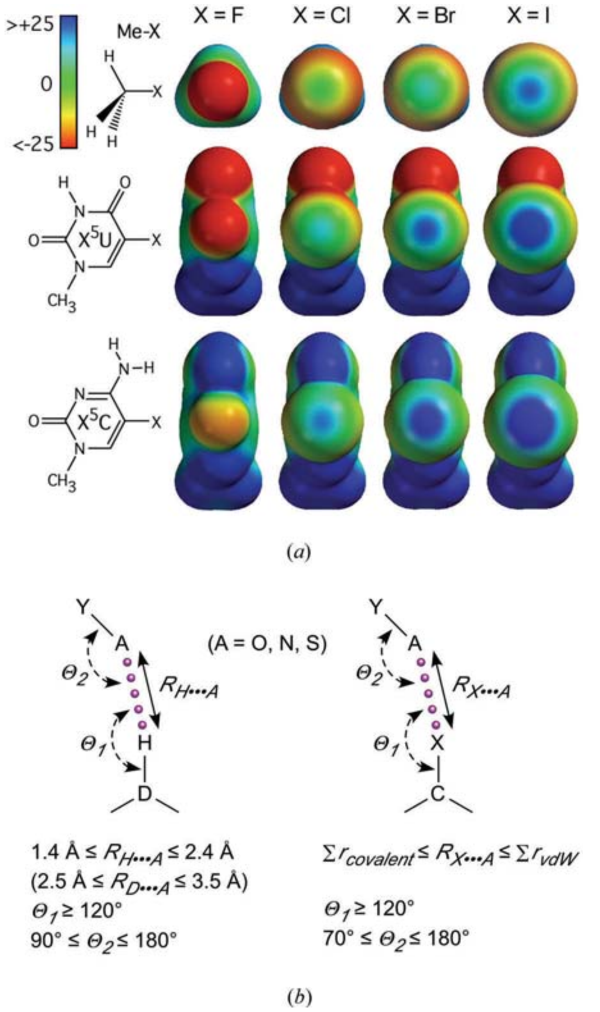


Figure 23.6.3.2
Properties of halogen bonds. (*a*) *Ab initio* calculations on model compounds (halomethane top, 5-halouracil middle, 5-halocytosine bottom) show an increase in polarizability of molecular halogens (X), with I > Br > Cl > F [adapted from Auffinger *et al.* (2004)]. The molecular surfaces are shown looking down the C–X bond, with electropositive and electronegative potentials shown in blue and red, respectively. (*b*) Comparison of the geometries of hydrogen bonds (left) and halogen bonds (right) [adapted from Voth & Ho (2007)]. In each case, the weak interactions are characterized by the distance between the donor (D) and acceptor (A) atoms ($R_{H\cdots A}$ and $R_{X\cdots A}$), which is longer than the sum of the covalent radii ($r_{\rm covalent}$) but shorter than the sum of the van der Waals radii ($r_{\rm vdW}$), and by the defined angles for the approach of the acceptor towards the donor (θ_1) and the donor towards the acceptor (θ_2).

The type of molecular interaction described by Hassel involves halogens serving as Lewis acids rather than as Lewis bases, and has seen a resurgence of interest in materials science. These interactions, initially called 'charge-transfer bonds' (Hassel, 1972), could, for example, draw the Br atoms of the Br_2 molecule to within 2.7 Å of the O atoms in dioxane (~0.5 Å shorter than $\sum r_{\rm vdW}$ of the interacting atoms). This would initially appear to be contrary to the expectation that electron-rich atoms should repel each other, but, as we shall see, halogens are not uniform spheres of electron density. In halogens larger than fluorine, the electron distributions are polarized such that an electropositive crown develops along the covalent (σ) C—X bond (Fig. 23.6.3.2*a*), and it is this 'σ hole' (Politzer, Murray & Lane, 2007) that accommodates the stabilizing electrostatic interactions between a molecular halogen and an electron-rich Lewis base (Fig. 23.6.3.2*b*). Such interactions are now referred to as halogen bonds (or X bonds) to acknowledge their similarities with the better known hydrogen bond (Metrangolo *et al.*, 2005; Metrangolo & Resnati, 2008). We can thus define the X-bond acceptor (by analogy with hydrogen-bond acceptors) as any electron-rich Lewis base, and the X-bond donor as the electropositive crown of the polarized halogen. Although they share similar acceptors (Metrangolo & Resnati, 2001) and, as we will discuss in greater detail, similar energies of interaction, X bonds are more directional than hydrogen bonds (Lommerse *et al.*, 1996).

The 'strength' of the X bond is expected to be dependent on the polarizability of the halogen donor, with I > Br > Cl in both polarizability and X-bonding potential (F atoms are not generally polarizable and therefore not typically considered X-bond donors) (Fig. 23.6.3.2). This is consistent with the understanding that the X bond is primarily an electrostatic interaction, although dispersion and charge transfer do contribute to some extent (Metrangolo *et al.*, 2005). Today, X bonds are exploited in chemistry to control molecular crystallization and supramolecular assembly (Metrangolo *et al.*, 2005; Metrangolo & Resnati, 2008; Politzer, Murray & Concha, 2007).

With this long history and a rekindled interest in applying X bonds for the design and synthesis of new materials (Fourmigué, 2009), it is surprising that this interaction has been largely overlooked in structural biology, but this could be attributed, until recently, to the very limited number of biological structures that included Group VII atoms. In 2004, we surveyed the occurrence of X bonds in the PDB and found 113 unique X-bond interactions in 66 protein–ligand complexes and six nucleic acid structures (Auffinger *et al.*, 2004). The results were dominated by X bonds to O atoms, with carbonyl O atoms of the peptide backbone (Figs. 23.6.1.1*a*,*b*) in protein–ligand complexes representing ~70% of these interactions. A survey of the 2008 PDB entries by Lu *et al.* (2009) found 258 X bonds to O-atom acceptors in the structures of protein–ligand complexes alone. Clearly, X-bonding interactions are important for ligand binding to proteins and this list of X-bonding interactions is expected to grow with the PDB, while becoming more prominent in the literature as the significance of this interaction becomes better recognized. A short Br$\cdots$O X bond is estimated to contribute to a more than 1000-fold increase in selectivity (equivalent to ~4 kcal mol^{-1} in energy; 1 kcal mol^{-1} = 4.184 kJ mol^{-1}) for the binding of an inhibitor by aldose reductase *versus* the related aldehyde reductase enzyme (Muzet *et al.*, 2003). In addition to oxygen, other electron-rich Lewis bases are also seen as X-bond acceptors, including the N atoms of various basic amino-acid side chains (Voth & Ho, 2007) and the S atoms of Cys or Met residues (Liu *et al.*, 2009). Much of the remainder of this discussion will be on the dominant interactions with O-atom acceptors of various types.

The geometry of X bonds provides an understanding of what is involved in the formation and stabilization of X bonds in biological systems. By analogy with the structure of X bonds between small organic molecules (Cody & Murray-Rust, 1984; Lommerse *et al.*, 1996; Ouvrard *et al.*, 2003), the geometry of biological X bonds is characterized by the angles of approach of the acceptor to the halogen donor (θ_1 relative to the C—X bond; Fig. 23.6.3.2*b*) and the halogen towards the base acceptor (θ_2; Fig. 23.6.3.2*b*) (Auffinger *et al.*, 2004). The θ_1 angle is essentially linear, consistent with this being an electrostatic interaction of the base acceptor, with the electropositive crown, or σ hole, being aligned along the C—X bond, as seen previously in organic complexes. The θ_2 angle is ~110–120°, suggesting that the X bonds involve the non-bonding electrons of the acceptor. However, an extended analysis showed that most X bonds in protein–ligand complexes are aligned perpendicular to the

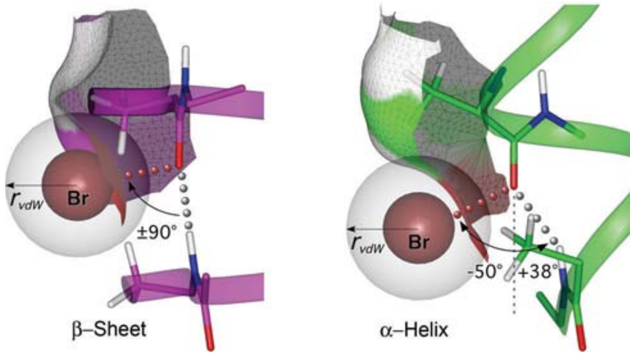


Figure 23.6.3.3
The orthogonal relationship between hydrogen bonds and X bonds that share common carbonyl O atoms as acceptors (Voth *et al.*, 2009), showing the approach of halogens from the ligands towards the carbonyl O atoms of the peptide bonds in ideal models of a β-sheet (left) and an α-helix (right) [adapted from Voth *et al.* (2009)]. In the β-sheet, the hydrogen bond to the shared O atom is oriented in the plane of the peptide bond, allowing the orthogonal X bond to approach from above or below and perpendicular to the plane. In the α-helix, the hydrogen bond is oriented ~40° from the plane and therefore the orthogonal relationship aligns the X bond ~50° from the peptide plane.

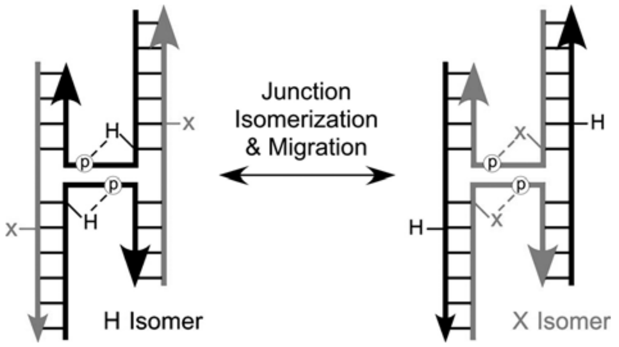


Figure 23.6.3.4
Competition assay to measure the energies of X bonds *versus* hydrogen bonds using a DNA junction (Voth *et al.*, 2007). The junction is constructed by annealing two hydrogen-bonded strands (black) and two X-bonded strands (grey). It is stabilized either by hydrogen bonds from the cytosine bases to the phosphates (labeled 'p') of the crossover DNA strand to form the H isomer (left), or by X bonds from a 5-halouracil at the equivalent position of the X-bonded strand to form the X isomer (right). The distribution of X *versus* H isomers in the crystal structure can be determined by analysing the distribution of halogens on the inside *versus* the outside strands of the junction, which, in turn, can be used to calculate the difference in energy between the X and hydrogen bonds at this position in the DNA junction (assuming that the crystal samples the population in the solution and that the crystal structures for the H and X isomers are nearly identical).

peptide plane, indicating that it is the π- rather than n-electron system that is preferred for X-bonding interactions (Auffinger *et al.*, 2004).

This preference for the π system in biological X bonds is not what would be predicted from surveys of small-molecule structures. In organic complexes, an X bond to a π system is seen rarely and only when there appears to be significant steric crowding that limits access to the X-bond acceptor atom (Ouvrard *et al.*, 2003). We have proposed the concept that X bonds in proteins are orthogonal molecular interactions to hydrogen bonds when they share a common carbonyl oxygen acceptor (Fig. 23.6.3.3) (Voth *et al.*, 2009), with orthogonality referring to both the perpendicular orientation of the shared X and hydrogen bonds and the relative thermodynamic independence of the two interactions. This model suggests that X bonds can be introduced in a well defined geometry to enhance the affinity at a ligand–protein interaction interface without disrupting the structural hydrogen bonds of the protein target. This is potentially a very powerful concept when X bonds are applied as a molecular tool for the rational design and synthesis of new inhibitors and drugs against enzyme targets (Metrangolo & Resnati, 2008).

How strong is a biological X bond? An X bond has been estimated, from *ab initio* calculations and crystallization of organic molecules, to be equivalent to a hydrogen bond in energy. Similar calculations on X bonds to peptide O atoms in protein–ligand complexes suggest that a hydrogen bond has approximately the same energy as an iodine X bond, with I > Br > Cl, as expected from their relative polarizabilities (Lu *et al.*, 2009). This is consistent with the observation that halogenated benzenes can form halogen bonds to the S atom of a Met residue in an artificial pocket engineered into a T4 lysozyme mutant (Liu *et al.*, 2009), but that the energies of these interactions are weak (approximately 0.5–0.7 kcal mol^{-1} more stabilizing than simple van der Waals interactions).

The energies of X bonds *versus* hydrogen bonds in a nucleic acid system have been estimated directly using a crystallographic assay that takes advantage of the conformational isomerization of a DNA Holliday junction (Grainger *et al.*, 1998; Lilley, 1999; McKinney *et al.*, 2003) and the ability of specific hydrogen and X bonds to the phosphate O atoms to stabilize the four-stranded complex (Eichman *et al.*, 2001; Hays *et al.*, 2005, 2003) (Fig. 23.6.3.4). From this set of competition studies, it was determined that a $\text{Br}\cdots\text{O}^{-1/2}$ type X bond is ~2–5 kcal mol^{-1} more stable than a hydrogen bond (Voth *et al.*, 2007). When extended to other halogens, the studies showed that I > Br > H > F (unpublished work). This competition assay also demonstrates that X bonds can be used to control DNA conformation (Voth *et al.*, 2007).

The energies determined for the $\text{C}{-}X\cdots\text{O}^{-1/2}$ X bonds in this DNA system (Voth *et al.*, 2007) are at least four to ten times greater than those estimated for the $\text{C}{-}X\cdots\text{S}$ X bonds in the T4 lysozyme–ligand studies (Liu *et al.*, 2009), and one might ask whether these discrepancies can be reconciled. We suggest that it is the nature of the Lewis-base acceptor that accounts for the dramatic difference, reflecting perhaps the difference between the base strength of an S atom in a thiol ether and that of the formally charged O atom of a phosphodiester. This hypothesis can be tested by comparing the energies of a neutral and a formally charged phosphate O atom. The stabilizing potential of a biological X bond apparently falls within a very broad range from near zero to a value >5 kcal mol^{-1} more stable than a standard hydrogen bond, depending on the polarizability of the donor halogen and the basicity of the acceptor. Thus, X bonds can be applied as a molecular tool that can be tuned energetically to control and drive macromolecular interactions, which can be exploited in synthetic biology for creating new nanomaterials from biological molecules.

Although oxygen, nitrogen and sulfur are the dominant X-bond acceptors, the side chains of aromatic amino acids can also serve as X-bond acceptors (Voth & Ho, 2007). In the specific examples available for ligand complexes with protein kinases, the $\text{C}{-}X$ bonds are oriented perpendicular to the plane and impinge on the centre of the aromatic ring (Fig. 23.6.1.1*b*), similar to hydrogen bonds to aromatic amino acids and cation–π interactions between positively charged and aromatic amino acids (Meyer *et al.*, 2003). The X-bond acceptor, in these cases, would be the delocalized π-electron system of the aromatic side chains. *Ab initio* calculations suggest that the electrostatic contribution to aromatic X bonds is very weak (around 1 kcal mol^{-1} at an optimum distance of 3.6 Å from a Br atom to the centre of an aromatic ring, or ~0.3 Å shorter than the predicted closest van der Waals approach; Voth & Ho, 2007), but hydrophobic effects may enhance the stabilizing potential of this type of interaction geometry.

23.6.3.2. Other halogen-specific electrostatic interactions in biological macromolecules

Specific X-bonding interactions between electropositive σ holes and Lewis-base acceptors are only one class of electrostatic interaction available to halogens. There are other classes that are variations on the same theme, where the halogen serves as a Lewis acid attracted to an electron-rich base, or where the halogen acts as the base attracted to an electrophilic acid. The most obvious example would be an electronegative F atom (which is not polarized) interacting with electropositive atoms. A very clear example of this type of electrostatic interaction is seen in the short contact (2.9 Å) between one F atom of the ZK-807834 inhibitor and the amide C atom of the Gln192 side chain in the ligand-binding site of the clotting factor Xa (Fig. 23.6.1.1*c*) (Adler *et al.*, 2000). We should note that, in addition to creating an electropositive crown, polarization of a halogen will also render the annulus around the σ hole more electronegative. This electronegative ring can serve as a Lewis base to interact with electropositive atoms and is observed in interactions of Cl, Br and I from ligands with the carbonyl C atoms, as opposed to the O atoms of the peptide backbone and various amino-acid side chains (Paulini *et al.*, 2005).

The distinction between X bonds and these other types of electrostatic interaction for any given halogen atom is dependent on the geometry of approach and the electrostatic nature of the atom approaching the halogen. An electronegative base approaching linearly to the C—X bond towards the σ hole would define an X bond, while an electropositive donor approaching perpendicular to the C—X bond would probably not. For example, a survey of short contacts between halogens and aromatic side chains (Saraogi *et al.*, 2003) showed that F atoms were the most commonly seen to interact with His, Phe, Trp and Tyr residues, although I, Br and Cl also show interactions with various geometries. The analysis indicates that fluorine interactions show a single broad distribution for distances from the halogen to the aromatic ring, while the distance distribution for chlorine appears to be bimodal, with one group at very short distances and another at about the van der Waals contact distance. A more complete analysis to correlate the distance–angle relationships may help to distinguish between aromatic X bonds to delocalized π systems and the opposing electrostatic interactions to the more electropositive carbon centres.

23.6.4. Concluding remarks

Halogens show a variety of different types of interaction with atoms in proteins and nucleic acids, and the particular type of interaction depends on both the properties of the halogen and what it is in contact with. It is this variability that may account for the distinct preference for certain halogens to interact with particular types of amino acids in proteins (Kortagere *et al.*, 2008). However, unlike most other interactions there is a strong directionality to molecular halogen interactions, particularly in biological systems, which may specify one type of interaction over another.

If we could fully understand the relationships between the geometries and energies of interactions, the types of electrostatic interaction described here could be exploited in biomolecular engineering in ways that are similar to, and extensions beyond, the principles used to design synthetic materials. We should stress, however, that our understanding of such interactions, particularly of biological X bonds, is still in its infancy. For example, Kraut *et al.* (2009) engineered halogenated phenylalanine residues into ketosteroid isomerase in order to determine whether an X bond could replace a functionally important hydrogen-bond interaction with an oxyanion hole, and found that the halogenated residues behaved similarly to a non-halogenated Phe residue at this functional position. However, close inspection of the model for the X–Phe mutant in this construct suggests that the geometry of approach of the halogen towards the oxyanion may not be the ideal for a particularly strong X bond. Thus, although it is tempting to apply such interactions in rational design strategies, we may simply not have enough understanding yet of the basic underlying principles to exploit these interactions effectively at this time. In another case, formation of the hairpin conformation that forms a kissing-loop structure in the RNA dimerization initiation site (DIS) from HIV-1 was shown to be facilitated by bromination of the uracil at the U3 position, compared with the bromination of uracils at other positions and the non-brominated RNA, which favour the duplex form (Ennifar *et al.*, 2007). The crystal structures of the brominated U3 *versus* other uracils provide no insight into why this particular halogen at this specific position should have such a profound effect on the conformational preference of the RNA.

Still, we would argue that a combination of experimental studies, coupled with *ab initio* and database analyses, will provide detailed understanding of the geometry–energy relationships for the various types of halogen interaction available to biological macromolecules. With this understanding, novel molecular tools can be developed to facilitate the rational design of new and more effective inhibitors against therapeutic targets, to manipulate the conformation and stability of protein and nucleic acid structures in the design of new nanomaterials, and to design new interaction interfaces for protein–protein and protein–nucleic acid complexes in synthetic biology.

References

Adler, M., Davey, D. D., Phillips, G. B., Kim, S. H., Jancarik, J., Rumennik, G., Light, D. R. & Whitlow, M. (2000). *Preparation, characterization, and the crystal structure of the inhibitor ZK-807834 (CI-1031) complexed with factor Xa. Biochemistry*, **39**, 12534–12542.

Auffinger, P., Hays, F. A., Westhof, E. & Ho, P. S. (2004). *Halogen bonds in biological molecules. Proc. Natl Acad. Sci. USA*, **101**, 16789–16794.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank. Nucleic Acids Res.* **28**, 235–242.

Bondi, A. (1964). *van der Waals volumes and radii. J. Phys. Chem.* **68**, 441–451.

Buss, I. H., Senthilmohan, R., Darlow, B. A., Mogridge, N., Kettle, A. J. & Winterbourn, C. C. (2003). *3-Chlorotyrosine as a marker of protein damage by myeloperoxidase in tracheal aspirates from preterm infants: association with adverse respiratory outcome. Pediatr. Res.* **53**, 455–462.

Cody, V. & Murray-Rust, P. (1984). *Iodine X(O, N, S) intermolecular contacts: models of thyroid hormone-protein binding interactions using information from the Cambridge crystallographic data files. J. Mol. Struct.* **112**, 189–199.

De Moliner, E., Brown, N. R. & Johnson, L. N. (2003). *Alternative binding modes of an inhibitor to two different kinases. Eur. J. Biochem.* **270**, 3174–3181.

Egner, U., Kratzschmar, J., Kreft, B., Pohlenz, H. D. & Schneider, M. (2005). *The target discovery process. ChemBioChem*, **6**, 468–479.

Eichman, B. F., Mooers, B. H. M., Alberti, M., Hearst, J. E. & Ho, P. S. (2001). *The crystal structures of psoralen cross-linked DNAs: drug dependent formation of Holliday junctions. J. Mol. Biol.* **301**, 15–26.

Ennifar, E., Bernacchi, S., Wolff, P. & Dumas, P. (2007). *Influence of C-5 halogenation of uridines on hairpin versus duplex RNA folding. RNA*, **13**, 1445–1452.

Fourmigué, M. (2009). *Halogen bonding: recent advances. Curr. Opin. Solid State Mater. Sci.* **13**, 36–45.

Grainger, R. J., Murchie, A. I. & Lilley, D. M. (1998). *Exchange between stacking conformers in a four-way DNA junction. Biochemistry*, **37**, 23–32.

Gribble, G. W. (2004). *Natural organohalogens: a new frontier for medicinal agents? J. Chem. Educ.* **81**, 1441–1449.

Hammett, L. P. (1937). *The effect of structure upon the reactions of organic compounds. Benzene derivatives. J. Am. Chem. Soc.* **59**, 96–103.

Hammill, J. T., Miyake-Stoner, S., Hazen, J. L., Jackson, J. C. & Mehl, R. A. (2007). *Preparation of site-specifically labeled fluorinated proteins for 19F-NMR structural characterization. Nat. Protoc.* **2**, 2601–2607.

Hansch, C. L. A. (1979). *Substituent Constants for Correlation Analysis in Chemistry and Biology*. New York: Wiley.

Hart, D. O., He, S., Chany, C. J. II, Withers, S. G., Sims, P. F., Sinnott, M. L. & Brumer, H. III (2000). *Identification of Asp-130 as the catalytic nucleophile in the main alpha-galactosidase from Phanerochaete chrysosporium, a family 27 glycosyl hydrolase. Biochemistry*, **39**, 9826–9836.

Hassel, O. (1972). *Structural aspects of interatomic charge-transfer bonding*. In *Nobel Lectures, Chemistry 1963–1970*. Amsterdam: Elsevier.

Hays, F. A., Teegarden, A., Jones, Z. J., Harms, M., Raup, D., Watson, J., Cavaliere, E. & Ho, P. S. (2005). *How sequence defines structure: a crystallographic map of DNA structure and conformation. Proc. Natl Acad. Sci. USA*, **102**, 7157–7162.

Hays, F. A., Vargason, J. M. & Ho, P. S. (2003). *Effect of sequence on the conformation of DNA Holliday junctions. Biochemistry*, **42**, 9586–9597.

Kortagere, S., Ekins, S. & Welsh, W. J. (2008). *Halogenated ligands and their interactions with amino acids: implications for structure–activity and structure–toxicity relationships. J. Mol. Graph. Model.* **27**, 170–177.

Kraut, D. A., Churchill, M. J., Dawson, P. E. & Herschlag, D. (2009). *Evaluating the potential for halogen bonding in the oxyanion hole of ketosteroid isomerase using unnatural amino acid mutagenesis. Am. Chem. Soc. Chem. Biol.* **4**, 269–273.

Lilley, D. M. J. (1999). *Oxford Handbook of Nucleic Acid Structure*, edited by S. Neidle, pp. 471–498. New York: Oxford University Press.

Liu, L., Baase, W. A. & Matthews, B. W. (2009). *Halogenated benzenes bound within a non-polar cavity in T4 lysozyme provide examples of I···S and I···Se halogen-bonding. J. Mol. Biol.* **385**, 595–605.

Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). *The nature and geometry of intramolecular interactions between halogens and oxygen or nitrogen. J. Am. Chem. Soc.* **118**, 3108–3116.

Lougheed, B., Ly, H. D., Wakarchuk, W. W. & Withers, S. G. (1999). *Glycosyl fluorides can function as substrates for nucleotide phosphosugar-dependent glycosyltransferases. J. Biol. Chem.* **274**, 37717–37722.

Lu, Y., Shi, T., Wang, Y., Yang, H., Yan, X., Luo, X., Jiang, H. & Zhu, W. (2009). *Halogen bonding – a novel interaction for rational drug design? J. Med. Chem.* **52**, 2854–2862.

Ly, H. D., Howard, S., Shum, K., He, S., Zhu, A. & Withers, S. G. (2000). *The synthesis, testing and use of 5-fluoro-alpha-D-galactosyl fluoride to trap an intermediate on green coffee bean alpha-galactosidase and identify the catalytic nucleophile. Carbohydr. Res.* **329**, 539–547.

McKinney, S. A., Déclais, A. C., Lilley, D. M. & Ha, T. (2003). *Structural dynamics of individual Holliday junctions. Nat. Struct. Biol.* **10**, 93–97.

Metrangolo, P., Neukirch, H., Pilati, T. & Resnati, G. (2005). *Halogen bonding based recognition processes: a world parallel to hydrogen bonding. Acc. Chem. Res.* **38**, 386–395.

Metrangolo, P. & Resnati, G. (2001). *Halogen bonding: a paradigm in supramolecular chemistry. Chemistry*, **7**, 2511–2519.

Metrangolo, P. & Resnati, G. (2008). *Chemistry. Halogen versus hydrogen. Science*, **321**, 918–919.

Meyer, E. A., Castellano, R. K. & Diederich, F. (2003). *Interactions with aromatic rings in chemical and biological recognition. Angew. Chem. Int. Ed. Engl.* **42**, 1210–1250.

Muzet, N., Guillot, B., Jelsch, C., Howard, E. & Lecomte, C. (2003). *Electrostatic complementarity in an aldose reductase complex from ultra-high-resolution crystallography and first-principles calculations. Proc. Natl Acad. Sci. USA*, **100**, 8742–8747.

Ouvrard, C., Le Questel, J.-Y., Berthelot, M. & Laurence, C. (2003). *Halogen-bond geometry: a crystallographic database investigation of dihalogen complexes. Acta Cryst.* B**59**, 512–526.

Paulini, R., Müller, K. & Diederich, F. (2005). *Orthogonal multipolar interactions in structural chemistry and biology. Angew. Chem. Int. Ed. Engl.* **44**, 1788–1805.

Pée, K. H. van & Unversucht, S. (2003). *Biological dehalogenation and halogenation reactions. Chemosphere*, **52**, 299–312.

Politzer, P., Murray, J. S. & Concha, M. C. (2007). *Halogen bonding and the design of new materials: organic bromides, chlorides and perhaps even fluorides as donors. J. Mol. Model.* **13**, 643–650.

Politzer, P., Murray, J. S. & Lane, P. (2007). *Sigma-hole bonding and hydrogen bonding: competitive interactions. Int. J. Quantum Chem.* **107**, 3046–3052.

Saraogi, I., Vijay, V., Das, S., Sekar, K. & Guru Row, T. (2003). *C—halogen···π interactions in proteins: a database study. Cryst. Eng.* **6**, 69–77.

Scott, M. E. & Viola, R. E. (1998). *The use of fluoro- and deoxy-substrate analogs to examine binding specificity and catalysis in the enzymes of the sorbitol pathway. Carbohydr. Res.* **313**, 247–253.

Shen, Z., Mitra, S. N., Wu, W., Chen, Y., Yang, Y., Qin, J. & Hazen, S. L. (2001). *Eosinophil peroxidase catalyzes bromination of free nucleosides and double-stranded DNA. Biochemistry*, **40**, 2041–2051.

Voth, A. R., Hays, F. A. & Ho, P. S. (2007). *Directing macromolecular conformation by halogen bonds. Proc. Natl Acad. Sci. USA*, **104**, 6188–6193.

Voth, A. R. & Ho, P. S. (2007). *The role of halogen bonding in inhibitor recognition and binding by protein kinases. Curr. Top. Med. Chem.* **7**, 1336–1348.

Voth, A. R., Khuu, P., Oishi, K. & Ho, P. S. (2009). *Halogen bonds as orthogonal molecular interactions to hydrogen bonds. Nat. Chem.* **1**, 74–79.

Wang, L., Xie, J. & Schultz, P. G. (2006). *Expanding the genetic code. Annu. Rev. Biophys. Biomol. Struct.* **35**, 225–249.

Whittington, D. A., Rosenzweig, A. C., Frederick, C. A. & Lippard, S. J. (2001). *Xenon and halogenated alkanes track putative substrate binding cavities in the soluble methane monooxygenase hydroxylase. Biochemistry*, **40**, 3476–3482.

Wojtczak, A., Cody, V., Luft, J. R. & Pangborn, W. (2001). *Structure of rat transthyretin (rTTR) complex with thyroxine at 2.5 Å resolution: first non-biased insight into thyroxine binding reveals different hormone orientation in two binding sites. Acta Cryst.* D**57**, 1061–1070.

Wu, W., Samoszuk, M. K., Comhair, S. A., Thomassen, M. J., Farver, C. F., Dweik, R. A., Kavuru, M. S., Erzurum, S. C. & Hazen, S. L. (2000). *Eosinophils generate brominating oxidants in allergen-induced asthma. J. Clin. Invest.* **105**, 1455–1463.

Xie, J., Wang, L., Wu, N., Brock, A., Spraggon, G. & Schultz, P. G. (2004). *The site-specific incorporation of p-iodo-L-phenylalanine into proteins for structure determination. Nat. Biotechnol.* **22**, 1297–1301.

International Tables for Crystallography (2012). Vol. F, Chapter 24.1, pp. 827–832.

PART 24. CRYSTALLOGRAPHIC DATABASES

Chapter 24.1. The Worldwide Protein Data Bank

H. M. Berman, K. Henrick, G. Kleywegt, H. Nakamura and J. Markley

24.1.1. Introduction

The Protein Data Bank (PDB) was established at Brookhaven National Laboratory (BNL) (Bernstein *et al.*, 1977) in 1971 as an archive for biological macromolecular crystal structures (see Chapter 24.1 in the first edition of this volume). It represents one of the earliest community-driven molecular biology data collections. In the beginning, the archive held seven structures, and with each passing year a handful more were deposited. In the 1980s, the number of deposited structures began to increase dramatically. This was due to technological improvements made in all aspects of the crystallographic process, the addition of structures determined by other methods and changes in community views about data sharing. By the early 1990s, many journals required a PDB ID for publication; now virtually all journals require the PDB deposition of not only coordinates, but also experimental data. As of 31 May 2011, the archive contained more than 73 000 structures.

The initial goal of the PDB was to archive structures determined by X-ray crystallography as submitted by the authors. Today, structures are deposited in the PDB that have been determined using X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and, more recently, cryo-electron microscopy (cryoEM). In addition to the structural biologists who deposit data, the international PDB community includes a diverse group of researchers in the biomedical sciences (biologists, chemists, physicists *etc.*), as well as educators and students at all levels. In order to ensure that the PDB remains a single uniform archive freely accessible *via* a collection of ftp tools, the worldwide PDB (wwPDB; http://www.wwpdb.org/) was formed in 2003 (Berman *et al.*, 2003, 2009) with an agreement between three deposition centres: the Research Collaboratory for Structural Bioinformatics (RCSB PDB) in the United States; the Protein Data Bank in Europe (PDBe; previously known as the Macromolecular Structure Database, MSD) at the European Bioinformatics Institute (EBI); and the PDB Japan (PDBj). The BioMagResBank (BMRB) in the United States joined in 2006. In this chapter, we describe the collection, validation, annotation and distribution of data by the wwPDB and provide a brief description of the services provided by the member sites.

24.1.2. Data acquisition and processing

24.1.2.1. Content of the data collected by the wwPDB

The PDB archive consists of entries containing three-dimensional Cartesian coordinates and information specific to the method of structure determination. Related experimental data include structure factors for X-ray experiments and chemical shifts and constraints derived from NMR experiments. NMR depositions in the PDB contain links to the BMRB archive of web-accessible constraints, chemical shifts and other data relevant to NMR structures (Ulrich *et al.*, 1989). In the near future, volumes for electron microscopy will become a part of the PDB archive. Table 24.1.2.1 lists the general information that the wwPDB collects for all structures as well as the information specific to X-ray, NMR and cryoEM experiments.

The definitions of the data items collected are in the PDB Exchange Dictionary (PDBx) (Westbrook, Henrick *et al.*, 2005), which is based on the mmCIF syntax. The mmCIF dictionary contains thousands of terms that define the macromolecular structure and the crystallographic experiment (Fitzgerald *et al.*, 2005). PDBx contains additional terms for NMR and cryoEM. The BMRB-developed NMR-STAR dictionary specifies additional data items linked to NMR structures. These terms were developed in collaboration with the depositors who are experts in these methods. Terms needed for tracking and other information management purposes are also included in PDBx.

24.1.2.2. Data deposition sites

Data are deposited to the PDB *via* one of the wwPDB member sites. Because it is critical that the final archive is kept uniform, the wwPDB works to make sure that the content and format of the final files, as well as the methods used to check them, are the same.

Each deposition to the PDB is represented by a PDB ID – a four character code. The PDB ID is assigned arbitrarily and is an immutable reference to the entry. PDB IDs are never reused and are used as a link between the structure and the literature reference that describes that structure. A unique and immutable integer tag identifies experimental NMR data that are processed and archived at the BMRB. Hyperlinks in the related PDB and BMRB files provide seamless access to all information.

The RCSB PDB (http://deposit.pdb.org/) and PDBj (http://pdbdep.protein.osaka-u.ac.jp/) use the program *ADIT* for data deposition and validation. PDBe processes and annotates data that are submitted *via AutoDep* (http://pdbe.org/deposit) (Tagari *et al.*, 2006). The wwPDB members are developing a single tool that will be used by all sites for deposition and annotation.

BMRB sites at Madison (http://deposit.bmrb.wisc.edu/bmrb-adit/) and Osaka-PDBj (http://nmradit.protein.osaka-u.ac.jp/bmrb-adit/) use *ADIT-NMR* to collect both coordinate and experimental data. NMR experimental data deposited at PDBe are forwarded to BMRB for further processing. NMR experimental data are processed by BMRB and the coordinate data by the other wwPDB sites.

Coordinates and experimental data files from all the sites are sent to the RCSB PDB for inclusion in the archive.

24.1.2.3. Validation and annotation

Validation refers to procedures for assessing the quality of the experimental data, the deposited atomic model and the fit of the model to the data. Annotation refers to the process of adding information to the entry (*e.g.*, cross-references to other biological

Copyright © 2012 International Union of Crystallography

Table 24.1.2.1
Content of data in the PDB

(*a*) *Content of all depositions*
Source – specifications such as genus, species, strain or variant of gene (cloned or synthetic); expression vector and host, or description of method of chemical synthesis
Sequence – full sequence of all macromolecular components
Chemical structure of cofactors and prosthetic groups
Names of all components of the structure
Broad description of the function and/or the composition of the structure
Literature citations for the structure submitted
Three-dimensional coordinates
Keywords and experimental method

(*b*) *Additional items for X-ray structure determinations*
Temperature factors and occupancies assigned to each atom
Crystallization conditions, including pH, temperature, solvents, salts, methods
Crystal data, including the unit-cell dimensions and space group
Presence of noncrystallographic symmetry
Data-collection information describing the methods used to collect the diffraction data including instrument, wavelength, temperature and processing programs
Data-collection statistics including data coverage, redundancy, $R_{\rm merge}$, $\langle I/\sigma(I)\rangle$, data above 1σ, 2σ, 3σ levels and resolution limits
Refinement information including (free) R factor, resolution limits, number of reflections, method of refinement, σ cutoff, geometry r.m.s.d., σ
Structure factors – h, k, l, $F_{\rm obs}$, $\sigma(F_{\rm obs})$, intensity, σ(intensity), flag for free R test

(*c*) *Additional items for NMR structure determinations*
For an ensemble, the model number for each coordinate set that is deposited and an indication whether one should be designated as a representative
Data-collection information describing the types of methods used, instrumentation, magnetic field strength, console, probe head, sample tube
Sample conditions, including solvent, macromolecule concentration ranges, concentration ranges of buffers, salts, antibacterial agents, other components, isotopic composition
Experimental conditions, including temperature, pH, pressure and oxidation state of structure determination, and estimates of uncertainties in these values
Non-covalent heterogeneity of sample, including self-aggregation, partial isotope exchange, conformational heterogeneity resulting in slow chemical exchange
Chemical heterogeneity of the sample (*e.g.*, evidence for deamidation or minor covalent species)
A list of NMR experiments used to determine the structure including those used to determine resonance assignments, NOE/ROE data, dynamical data, scalar coupling constants, and those used to infer hydrogen bonds and bound ligands. The relationship of these experiments to the constraint files are given explicitly
Constraint files used to derive the structure as described in Task Force recommendations
Links (where available) to associated files at BMRB containing chemical shift, filtered constraint and other NMR data

(*d*) *Additional data items for cryo-electron microscopy*
Sample preparation, including aggregation state, concentration, buffer, pH, sample support and description of vitrification procedure
3D reconstruction procedure, including method, nominal and actual pixel size, resolution, contrast-transfer-function correction method, magnification calibration and text description of the reconstruction procedure
3D model fitting procedure, including fitting-procedure type and program used, PDB IDs of models used
3D model refinement, including whether refinement is done in real space or reciprocal space, type of protocol, refinement target criteria
If sample is a 2D or 3D array, symmetry and repeat parameters
If the entry contains a virus, details of the virus host including type, species and growth cell, as well as details of the virus including type, isolate and International Committee on Taxonomy of Viruses ID

data resources). The wwPDB partners have convened three Validation Task Forces (VTFs) to advise them on how best to validate X-ray, NMR and cryoEM structures. Their recommendations will be implemented as part of the joint deposition and annotation system currently under development.

At present, annotation involves the following components:

Covalent bond distances and angles: Proteins are compared against standard geometry values from Engh & Huber (1991); nucleic acid bases are compared against standard values from Clowney *et al.* (1996); sugars and phosphates are compared against standard values from Gelbin *et al.* (1996).

Stereochemical validation: All chiral centres of proteins and nucleic acids are checked for correct stereochemistry.

Atom nomenclature: The nomenclature of all atoms is checked for compliance with IUPAC standards (IUPAC–IUB Joint Commission on Biochemical Nomenclature, 1983; Markley *et al.*, 1998) and adjusted if necessary.

Close contacts: The distances between all atoms within the asymmetric unit of crystal structures and the unique molecule of NMR structures are calculated. For crystal structures, contacts between symmetry-related molecules are checked as well.

Ligand and atom nomenclature: Residue and atom nomenclature are compared against a standard dictionary (http://www.wwpdb.org/ccd.html) for all ligands as well as standard residues and bases. Unrecognized ligand groups are flagged and any discrepancies in known ligands are listed as extra or missing atoms. New ligands are added to the dictionary as they are deposited.

Sequence comparison: The sequence provided by the depositor is compared against the sequence derived from the coordinate records. This information is displayed in a table where any differences or missing residues are annotated. During the annotation process the sequence database references provided by the author are checked for accuracy. If no reference is given, a sequence search against UniProtKB (The UniProt Consortium, 2011) is carried out to find the best match. Any conflict between the depositor's sequence and the sequence derived from the coordinate records is further resolved and annotated by comparison with other sequence databases as needed. Once an entry has been released, the cross-references and mappings to sequence and other databases are maintained and kept up to date in the separate SIFTS resource (http://pdbe.org/sifts), a joint project of PDBe and UniProt.

Distant waters: The distances between all water oxygen atoms and all polar atoms (oxygen and nitrogen) of the macromolecules, ligands and solvent in the asymmetric unit are calculated. Distant solvent atoms are repositioned using crystallographic symmetry such that they fall within the solvation sphere of the macromolecule.

Geometry: The torsion-angle distributions and peptide-bond deviations from *cis* and *trans* conformation are checked. Derived data, such as site records, helix and sheet records for secondary structure, are annotated.

Checks of NMR data by the BMRB: NMR constraints are checked for consistency with the three-dimensional structure and atom nomenclature, chemical shifts are checked for possible referencing errors and outliers.

In almost all cases, serious errors detected by these checks are corrected through annotation and correspondence with the authors.

The wwPDB continuously reviews its annotation methods and will continue to integrate new procedures as they become available and are accepted as community standards.

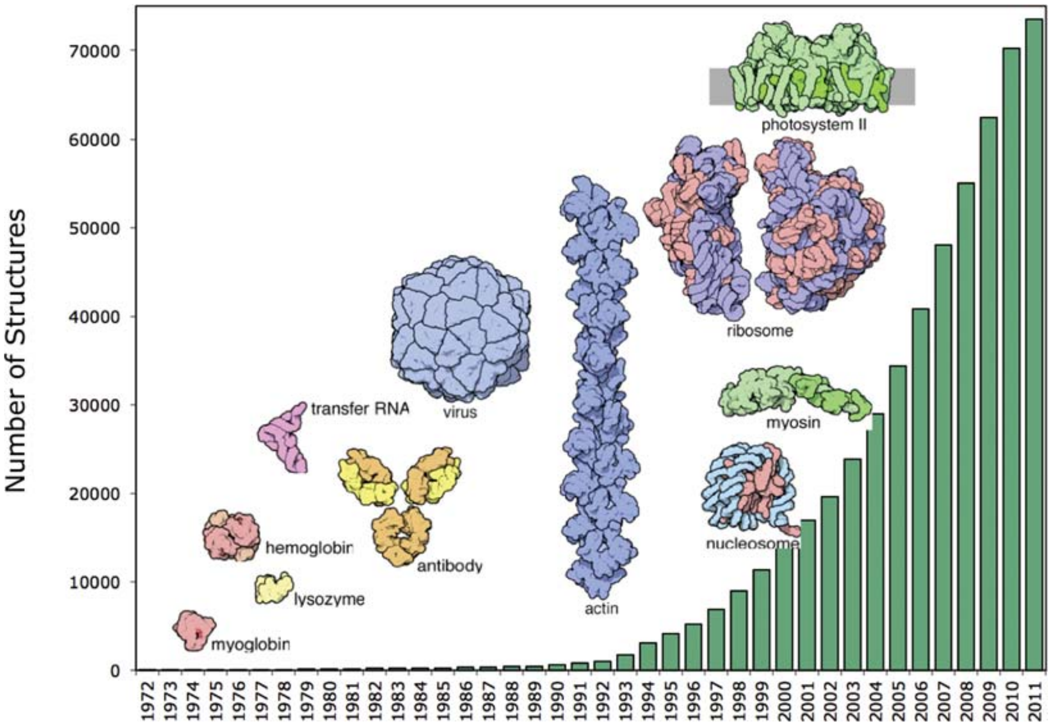


Figure 24.1.2.1
Growth chart of the PDB showing the total number of structures available in the PDB archive per year (as of 31 May 2011) and highlighting example structures from different time periods.

24.1.2.4. Data processing statistics

Fig. 24.1.2.1 shows the growth of PDB data since the archive began, indicating how the complexity of structures released into the archive has increased over time.

As of 31 May 2011, the PDB contains 73 503 publicly accessible structures; of these entries, 64 043 were determined by X-ray methods, 8902 by NMR and 370 by electron microscopy (Fig. 24.1.2.2).

24.1.2.5. Data uniformity

A key goal of the wwPDB is to make the archive as consistent and error-free as possible. Before files are released, all new depositions are reviewed by annotators. Errors found subsequent to release by authors and PDB users are addressed as rapidly as possible and the entries are revised.

The wwPDB member sites collaborate to review all the data in the PDB archive (Henrick *et al.*, 2008; Lawson *et al.*, 2008). In particular, the monomer components and ligands were examined to construct a dictionary that includes the stereochemistry, nomenclature, ideal coordinates, model coordinates, and SMILES and InChI representations. The Chemical Component Dictionary (http://www.wwpdb.org/ccd.html) contains the definitions for all components found in the PDB archive. All instances of these components have been matched against this dictionary and, where necessary, corrections have been made to the PDB files. The sequences, citations and some experimental data items were also corrected. More recently, the archive been reviewed and remediated with the objective of tackling complex problems, including the representation of biological assemblies, residual B factors, peptide inhibitors and antibiotics, and entries in nonstandard crystal frames. Documentation about ongoing remediation efforts is available from http://wwpdb.org.

24.1.3. Data access

24.1.3.1. ftp

The 'PDB archive' is the collection of flat files maintained in three formats (Westbrook & Fitzgerald, 2009) (http://www.wwpdb.org/docs.html): the legacy PDB file format (Bernstein *et al.*, 1977); the PDB exchange format (PDBx) that follows the mmCIF syntax (Fitzgerald *et al.*, 2005, Westbrook, Henrick *et al.*, 2005) (http://mmcif.pdb.org); and the PDBML/XML format (Westbrook, Ito *et al.*, 2005) that is a direct translation of the PDB exchange format. In addition to coordinate data, the archive contains experimental data (structure factors, NMR restraints and chemical shifts). Each wwPDB site distributes the same PDB archive *via* ftp. The archive is updated weekly.

Time-stamped snapshots of the PDB archive are added each year to ftp://snapshots.wwpdb.org. They provide a frozen copy of the archive as it appeared at that time for research and historical purposes. Scripts are available to download all, or part, of a snapshot automatically.

24.1.3.2. Websites

In addition to providing access to the PDB archive, each wwPDB site offers services and resources that provide different views and analyses of the structural data contained within the PDB archive.

24.1.3.2.1. RCSB PDB

The RCSB PDB (http://www.pdb.org) (Berman *et al.*, 2000) provides a website with several functionalities. In addition to downloading files in all formats, a variety of search capabilities are offered. These include searching by PDB ID, author or keyword; searching within a relational database that is populated by the data in the PDBx/PDBML files; or browsing characteristics that have been integrated from external resources (Deshpande *et al.*, 2005). These include biological process, cellular

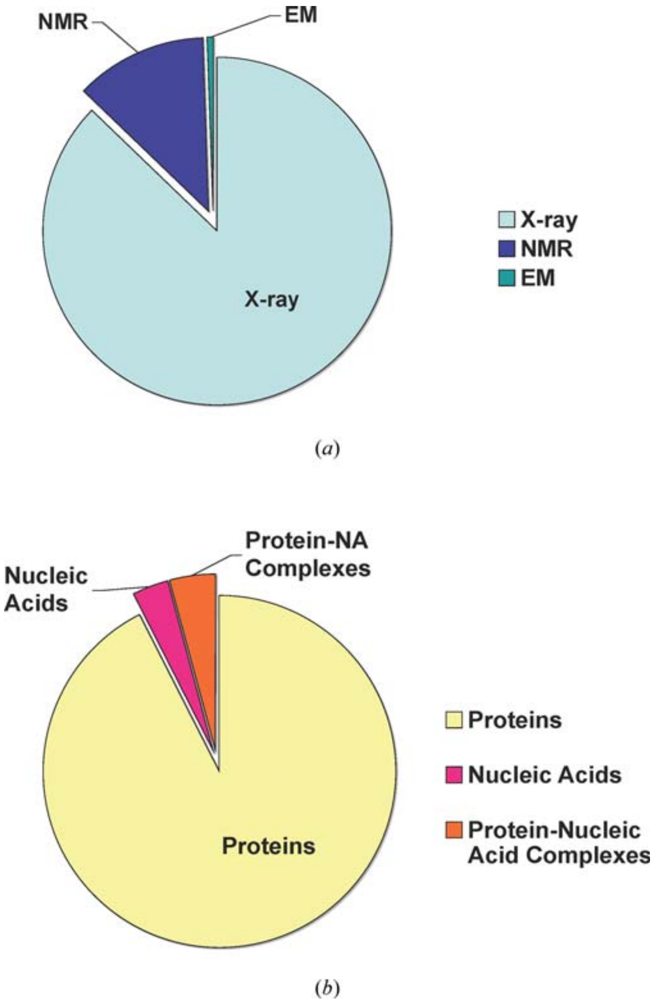


Figure 24.1.2.2
(*a*) Distribution of structures by experimental type (as of 31 May 2011). As of this date, the PDB has 73 503 publicly accessible structures; of these entries, 64 043 were determined by X-ray methods, 8902 by NMR and 370 by electron microscopy. (*b*) Distribution of structures by structure type (as of 31 May 2011). Of the currently available entries, 68 048 are proteins, 2257 are nucleic acids and 3159 are protein/nucleic acid complexes; 39 are other types of molecules.

component and molecular function as defined by gene ontology (The Gene Ontology Consortium, 2000); enzyme classification; medical subject headings (MeSH terms); source organism as defined by the NCBI taxonomy; gene location from Entrez (Wheeler *et al.*, 2006); and folds as classified by SCOP (Murzin *et al.*, 1995) and CATH (Orengo *et al.*, 1997). The structure summary report for each structure is itself searchable. A variety of tabular reports can be made for groups of structures with information about the experiment, the chemistry and biology, and citations. Static pictures and interactive graphics are provided for each structure and protein–ligand interaction from the website. Some of these interactive views use the molecular biology toolkit (*MBT*, http://mbt.sdsc.edu; Moreland *et al.*, 2005). A variety of summary statistics are available as histograms and in tabular form. Advanced searches can be performed on the structure data in combination with the data integrated from over 30 different external sources. Multimedia tutorials are available describing many other features.

A service called MyPDB notifies users *via* email when the PDB releases structures that match customized queries stored at the RCSB PDB site. For application developers, a variety of web services are available to directly access subsets of data and RCSB PDB resources.

Also published at the RCSB PDB site is the *Molecule of the Month*; authored and illustrated by David S. Goodsell, this feature describes a molecule in such a way that non-specialists can learn about the relationship between the structure and function of a particular molecular system.

24.1.3.2.2. PDBj

PDBj (http://www.pdbj.org) maintains a variety of services for querying, displaying and analysing PDB data (Standley *et al.*, 2008). *Structure Navigator* can be used to locate PDB entries that are similar to a query structure or PDB ID. Structural similarity is defined by the newly developed score used in the stand-alone program *ASH* (Alignment of Structural Homologs), which can be run or downloaded from http://www.pdbj.org/ASH/. *ASH* shows high sensitivity and selectivity in recognizing CATH or SCOP domains, without requiring large amounts of CPU (Standley *et al.*, 2004, 2005, 2007). *Sequence Navigator* is a *BLAST* interface to PDBj. The input to *Sequence Navigator* can be a PDB ID or an amino-acid sequence. *GIRAF* (Kinjo & Nakamura, 2009) is an efficient search method for the local atomic alignments of possible ligand binding sites and protein–protein interfaces (http://giraf.pdbj.org). It is not constrained by sequence homology, sequence order or protein fold, and hence is applicable to the cases where no structural homologues can be found to find clues about the possible function of a protein.

Most of the query services described above can be executed as web services using the Simple Object Application Protocol (SOAP), REST or through web browsers. *jV* version 3 (formerly known as *PDBjViewer*) is a molecular-graphics program to display proteins and nucleic acids (Kinoshita & Nakamura, 2004). eProtS, the Encyclopedia of Protein Structures, is an educational resource for learning about biological functions and structural characteristics of protein molecules. eProtS contains explanations for a few hundred protein molecules that are particularly interesting. eProtS is implemented as a Wiki so that motivated structural biologists can contribute articles for the proteins whose structures they have solved. Two derived databases are also available from PDBj. ProMode is a database of normal mode analyses (NMA) of proteins. eF-site (electrostatic surface of Functional site) is a database for molecular surfaces of protein functional sites, displaying the electrostatic potentials and hydrophobic properties together on the Connolly surfaces of the active sites, for analyses of the molecular-recognition mechanisms. eF-seek is a service to search for similar molecular surfaces among known active sites.

24.1.3.2.3. PDBe

PDBe (http://pdbe.org) (Velankar *et al.*, 2010, 2011) aims to bring structure to biology, *i.e.*, to become an integrated resource of structure-related data that provides relevant services, information and tools tailored to the needs of expert and non-expert users alike (Velankar & Kleywegt, 2011). Besides offering basic and advanced search and retrieval systems, PDBe focuses on advanced services, ligand-related resources, integration with other biological data resources, validation and enriched presentation of experimental data. Advanced PDBe services include *PDBeMotif* (http://pdbe.org/motif), a service for analysing detailed molecular interactions and correlating them with sequence or structure patterns (Golovin & Henrick, 2008); *PDBeFold* (http://pdbe.org/fold), a powerful interactive structure-alignment tool (Krissinel & Henrick, 2004); *PDBePISA* (http://pdbe.org/pisa), a quaternary-structure prediction and analysis service (Krissinel & Henrick, 2007); and *PDBeXplore* (http://pdbe.org/browse), a tool that allows browsing and analysis of the structural archive based on familiar chemical and biological classification systems such EC, CATH and Pfam (Velankar & Kleywegt, 2011).

Together with UniProt, PDBe also maintains the SIFTS resource (http://pdbe.org/sifts), which provides entry and residue-level mapping for proteins that are found in PDB entries, as well as cross-reference information to resources such as GO, Pfam, InterPro, CATH and SCOP (Velankar *et al.*, 2005). SIFTS data are used by many of the world's major structural bioinformatics resources (*e.g.*, RCSB PDB, Pfam, UniProt, SCOP) as a source of linking data about structure, sequence and function for PDB entries.

The PDBe website is also a rich source of meta-information about the PDB (statistics, highlights), educational material and Quips (short interactive articles and tutorials about 'quite interesting PDB structures'). PDBe further provides tools that can be included in other websites ('widgets') and that provide up-to-date information. Examples of these are PDBprints (http://pdbe.org/prints) and PDBportfolio (http://pdbe.org/portfolio), which provide key information and images, respectively, about individual PDB entries, and UniPDB (http://pdbe.org/unipdb), which shows the coverage of UniProt sequences by PDB entries. Finally, PDBe provides specialized information and services for the NMR (http://pdbe.org/nmr) and cryoEM (http://pdbe.org/em) communities.

24.1.3.2.4. BMRB

The BMRB (http://www.bmrb.wisc.edu) (Ulrich *et al.*, 1989) collects and disseminates experimental NMR data and joined the wwPDB in 2006 (Markley *et al.*, 2008). The BMRB archive can be searched by database ID, keywords, author, molecule name, sequence, sample conditions or other characteristics. The website contains statistics and resources for the NMR community. BMRB archives data connected with an NMR structure as recommended by the wwPDB NMR Task Force for structural-biology depositions: structural constraints, assigned chemical shifts, and NOESY spectra and/or peak lists. BMRB also supports more comprehensive NMR data depositions by members of the structural-genomics community, which may include additional primary (time-domain) spectra and intermediate results.

The BMRB provides software and tools for validating NMR data and associated structures prior to deposition and tools for visualizing various types of NMR data with or without an associated 3D structure. The BMRB is responsible for processing constraint files and for providing constraints that are consistent with structural models and converted to machine-readable NMR-STAR format to the PDB archive. Filtered restraints, further processed for use by test calculation protocols and longitudinal analyses and validations, are available online (Doreleijers *et al.*, 2005).

24.1.4. Future

Structural biology is a rapidly evolving field that constantly provides new challenges to the collection, curation and distribution of macromolecular structure data. Since 1998, the rate of depositions has tripled and the total holdings have increased by a factor of more than eight. The structures have increased in complexity, with many molecular machines now part of the archive. Data-harvesting tools have made it possible to collect and include more information about each structure. The challenge is to be able to continue to collect data rapidly while maintaining quality and uniformity across the archive.

The maintenance and further development of the PDB archive is a worldwide effort. The willingness of the global community to share ideas, software and data makes it a unique resource for biological research.

The RCSB PDB is supported by the National Science Foundation, the National Institute of General Medical Sciences, the Department of Energy, the National Library of Medicine, the National Cancer Institute, the National Institute of Neurological Disorders and Stroke, and the National Institute of Diabetes and Digestive and Kidney Diseases. Funding for PDBe is provided by EMBL–EBI, the Wellcome Trust, NIH, BBSRC and the European Union. Some earlier developments at PDBe were funded by the MRC and CCP4. PDBj is supported by the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, and the Ministry of Education, Culture, Sports, Science and Technology. The BMRB is supported by the National Library of Medicine.

This chapter is an updated version of Berman *et al.* (2009). Reprinted with permission of John Wiley & Sons, Inc.

References

Berman, H., Henrick, K. & Nakamura, H. (2003). *Announcing the worldwide Protein Data Bank*. *Nat. Struct. Biol.* **10**, 980.

Berman, H. M., Henrick, K., Nakamura, H. & Markley, J. L. (2009). *Structural Bioinformatics*, 2nd ed., edited by J. Gu & P. E. Bourne, pp. 293–303. Hoboken: John Wiley & Sons, Inc.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J. Mol. Biol.* **112**, 535–542.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: nitrogenous bases*. *J. Am. Chem. Soc.* **118**, 509–518.

Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z., Green, R. K., Flippen-Anderson, J. L., Westbrook, J., Berman, H. M. & Bourne, P. E. (2005). *The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema*. *Nucleic Acids Res.* **33**, D233–D237.

Doreleijers, J. F., Nederveen, A. J., Vranken, W., Lin, J., Bonvin, A. M., Kaptein, R., Markley, J. L. & Ulrich, E. L. (2005). *BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures*. *J. Biomol. NMR*, **32**, 1–12.

Engh, R. A. & Huber, R. (1991). *Accurate bond and angle parameters for X-ray protein structure refinement*. *Acta Cryst.* A**47**, 392–400.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, pp. 295–443. Dordrecht: Springer.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents*. *J. Am. Chem. Soc.* **118**, 519–528.

Golovin, A. & Henrick, K. (2008). *Msdmotif: exploring protein sites and motifs*. *BMC Bioinformatics*, **9**, 312.

Henrick, K., Feng, Z., Bluhm, W. F., Dimitropoulos, D., Doreleijers, J. F., Dutta, S., Flippen-Anderson, J. L., Ionides, J., Kamada, C., Krissinel, E., Lawson, C. L., Markley, J. L., Nakamura, H., Newman, R., Shimizu, Y., Swaminathan, J., Velankar, S., Ory, J., Ulrich, E. L., Vranken, W., Westbrook, J., Yamashita, R., Yang, H., Young, J., Yousufuddin, M. &

Berman, H. M. (2008). *Remediation of the Protein Data Bank archive*. *Nucleic Acids Res.* **36**, D426–D433.

IUPAC–IUB Joint Commission on Biochemical Nomenclature (1983). *Abbreviations and symbols for the description of conformations of polynucleotide chains*. *Eur. J. Biochem.* **131**, 9–15.

Kinjo, A. R. & Nakamura, H. (2009). *Comprehensive structural classification of ligand-binding motifs in proteins*. *Structure*, **17**, 234–246.

Kinoshita, K. & Nakamura, H. (2004). *eF-site and PDBjViewer: database and viewer for protein functional sites*. *Bioinformatics*, **20**, 1329–1330.

Krissinel, E. & Henrick, K. (2004). *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions*. *Acta Cryst.* D**60**, 2256–2268.

Krissinel, E. & Henrick, K. (2007). *Inference of macromolecular assemblies from crystalline state*. *J. Mol. Biol.* **372**, 774–797.

Lawson, C. L., Dutta, S., Westbrook, J. D., Henrick, K. & Berman, H. M. (2008). *Representation of viruses in the remediated PDB archive*. *Acta Cryst.* D**64**, 874–882.

Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E. & Wüthrich, K. (1998). *Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC–IUBMB–IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy*. *J. Biomol. NMR*, **12**, 1–23.

Markley, J. L., Ulrich, E. L., Berman, H. M., Henrick, K., Nakamura, H. & Akutsu, H. (2008). *BioMagResBank (BMRB) as a partner in the Worldwide Protein Data Bank (wwPDB): new policies affecting biomolecular NMR depositions*. *J. Biomol. NMR*, **40**, 153–155.

Moreland, J. L., Gramada, A., Buzko, O. V., Zhang, Q. & Bourne, P. E. (2005). *The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications*. *BMC Bioinformatics*, **6**, 21.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *SCOP: a structural classification of proteins database for the investigation of sequences and structures*. *J. Mol. Biol.* **247**, 536–540.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *CATH – a hierarchic classification of protein domain structures*. *Structure*, **5**, 1093–1108.

Standley, D. M., Kinjo, A. R., Kinoshita, K. & Nakamura, H. (2008). *Protein structure databases with new web services for structural biology and biomedical research*. *Brief. Bioinform.* **9**, 276–285.

Standley, D. M., Toh, H. & Nakamura, H. (2004). *Detecting local structural similarity in proteins by maximizing number of equivalent residues*. *Proteins*, **57**, 381–391.

Standley, D. M., Toh, H. & Nakamura, H. (2005). *Gash: An improved algorithm for maximizing the number of equivalent residues between two protein structures*. *BMC Bioinformatics*, **6**, 221.

Standley, D. M., Toh, H. & Nakamura, H. (2007). *Ash structure alignment package: sensitivity and selectivity in domain classification*. *BMC Bioinformatics*, **8**, 116.

Tagari, M., Tate, J., Swaminathan, G. J., Newman, R., Naim, A., Vranken, W., Kapopoulou, A., Hussain, A., Fillon, J., Henrick, K. & Velankar, S. (2006). *E-MSD: improving data deposition and structure quality*. *Nucleic Acids Res.* **34**, D287–D290.

The Gene Ontology Consortium (2000). *Gene ontology: tool for the unification of biology*. *Nat. Genet.* **25**, 25–29.

The UniProt Consortium (2011). *Ongoing and future developments at the Universal Protein Resource*. *Nucleic Acids Res.* **39**, D214–D219.

Ulrich, E. L., Markley, J. L. & Kyogoku, Y. (1989). *Creation of a nuclear magnetic resonance data repository and literature database*. *Protein Seq. Data Anal.* **2**, 23–37.

Velankar, S., Alhroub, Y., Alili, A., Best, C., Boutselakis, H. C., Caboche, S., Conroy, M. J., Dana, J. M., van Ginkel, G., Golovin, A., Gore, S. P., Gutmanas, A., Haslam, P., Hirshberg, M., John, M., Lagerstedt, I., Mir, S., Newman, L. E., Oldfield, T. J., Penkett, C. J., Pineda-Castillo, J., Rinaldi, L., Sahni, G., Sawka, G., Sen, S., Slowley, R., Sousa da Silva, A. W., Suarez-Uruena, A., Swaminathan, G. J., Symmons, M. F., Vranken, W. F., Wainwright, M. & Kleywegt, G. J. (2011). *PDBe: Protein Data Bank in Europe*. *Nucleic Acids Res.* **39**, D402–D410.

Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa da Silva, A. W., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., Krissinel, E. B., Newman, R., Oldfield, T., Pajon, A., Penkett, C. J., Pineda-Castillo, J., Sahni, G., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, J., van Ginkel, G., Vranken, W. F., Henrick, K. & Kleywegt, G. J. (2010). *PDBe: Protein Data Bank in Europe*. *Nucleic Acids Res.* **38**, D308–D317.

Velankar, S. & Kleywegt, G. J. (2011). *The Protein Data Bank in Europe (PDBe): bringing structure to biology*. *Acta Cryst.* D**67**, 324–330.

Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. & Henrick, K. (2005). *E-MSD: an integrated data resource for bioinformatics*. *Nucleic Acids Res.* **33**, D262–D265.

Westbrook, J., Henrick, K., Ulrich, E. L. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, pp. 195–198. Dordrecht: Springer.

Westbrook, J., Ito, N., Nakamura, H., Henrick, K. & Berman, H. M. (2005). *PDBML: the representation of archival macromolecular structure data in XML*. *Bioinformatics*, **21**, 988–992.

Westbrook, J. D. & Fitzgerald, P. M. D. (2009). *Structural Bioinformatics*, 2nd ed., edited by P. E. Bourne & J. Gu, pp. 271–291. Hoboken: John Wiley & Sons, Inc.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Helmberg, W., Kapustin, Y., Kenton, D. L., Khovayko, O., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Pruitt, K. D., Schuler, G. D., Schriml, L. M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T. O., Tatusov, R., Tatusova, T. A., Wagner, L. & Yaschenko, E. (2006). *Database resources of the National Center for Biotechnology Information*. *Nucleic Acids Res.* **34**, D173–D180.

International Tables for Crystallography (2012). Vol. F, Chapter 24.2, pp. 833–837.

Chapter 24.2. The Nucleic Acid Database

B. Schneider, J. de la Cruz, S. Dutta, Z. Feng, L. Chen, J. Westbrook, H. Yang, J. Young, C. Zardecki and H. M. Berman

24.2.1. Introduction

The Nucleic Acid Database (NDB) (Berman *et al.*, 1992) was established in 1991 as a resource for specialists in the field of nucleic acid structure. Over the years, the NDB has developed generalized software for processing, archiving, querying and distributing structural data for nucleic acid-containing structures. The core of the NDB has been its relational database of nucleic acid-containing crystal structures; it also incorporates structures derived by nuclear magnetic resonance spectroscopy (NMR). Recognizing the importance of a standard data representation in building a database, the NDB became an active participant in the mmCIF project (Fitzgerald *et al.*, 2005*a*,*b*) and was the test-bed for this format. With a foundation of well curated data, the NDB created a searchable relational database of primary and derivative data with very rich query and reporting capabilities. This robust database was unique in that it allowed researchers to perform comparative analyses of nucleic acid-containing structures according to the many attributes stored in the database.

In 1992, the NDB assumed responsibility for processing all nucleic acid crystal structures that were deposited into the Protein Data Bank (PDB) (Berman *et al.*, 2003, Bernstein *et al.*, 1977); it was the direct deposit site for nucleic acid structures from 1996 to 1998. In order to meet data-processing requirements, the NDB created the first validation software package for nucleic acids (Westbrook *et al.*, 2003), which is still used today. The NDB continues to provide a high level of information about nucleic acids and serves as a speciality database for its community of researchers. When the NDB began, the world of nucleic acid structures consisted of DNA and RNA oligonucleotides, a few protein–DNA complexes and some tRNA structures. Annotation of structural features was performed manually, and structures were easily classified into a few known molecular architectures by visual inspection. However, in the last several years a whole new universe of nucleic acid structures has emerged. There are many ribozyme structures and many different types of protein–nucleic acid complexes are represented. The additions of ribosomal structures to the archive have increased the number of nucleotide residues resident in the NDB several fold (Moore, 2001) since they emerged in 2000 (Fig. 24.2.1.1).

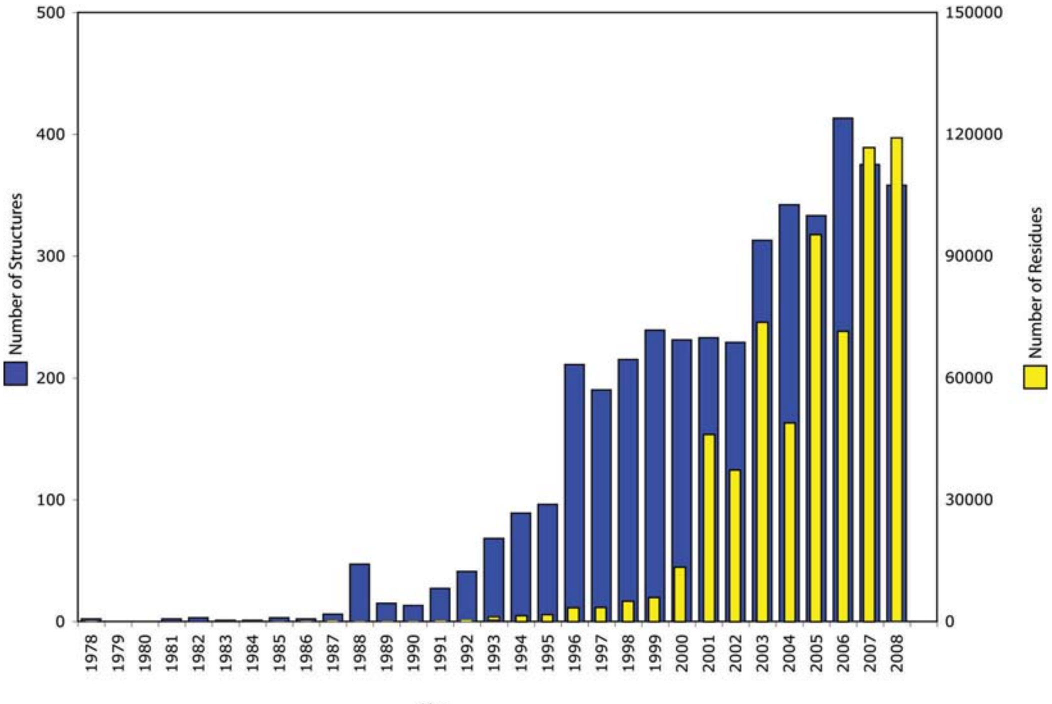


Figure 24.2.1.1
Growth of the NDB. The number of structures available released per year in the NDB is indicated in blue (scale on the left) and the number of nucleotides in these structures in yellow (scale on the right).

24.2.2. Data processing and validation

The NDB created a robust data-processing system to produce high-quality data that are readily loaded into a database. The full capability of this system was demonstrated by the successful processing of ribosomal subunits, which are very large and complex structures.

Early on, the NDB adopted the Macromolecular Crystallographic Information File (mmCIF; Bourne *et al.*, 1997; Fitzgerald *et al.*, 2005*a*,*b*; Westbrook *et al.*, 2005) as its data standard. This format has several advantages from the point of view of building a database: (1) the definitions for the data items are based on a comprehensive dictionary of crystallographic terminology and molecular structure description; (2) it is self-defining; and (3) the syntax contains explicit rules that further define the characteristics of the data items, particularly the relationships between data items (Westbrook & Bourne, 2000). The latter feature is important because it allows for rigorous checking of the data.

The tools first developed by the NDB project are used by the Research Collaboratory for Structural Bioinformatics PDB (Berman *et al.*, 2000) and the Protein Data Bank Japan (both members of the Worldwide PDB, http://www.wwpdb.org/) for processing both proteins and nucleic acids. The validation tool *NUCheck* verifies valence geometry, torsion angles, intermolecular contacts, and the chiral centres of the sugars and phosphates for nucleic acids (Westbrook *et al.*, 2003). The

Copyright © 2012 International Union of Crystallography

Table 24.2.3.1
The information content of the NDB

(*a*) *Identity of depositions, processing information*
Unique codes – NDB, PDB codes
Processing information – deposition, release dates

(*b*) *Primary experimental information stored in the NDB*
Chemical description – the chemical composition of the deposition (content of all chemical constituents of the deposition), detailed description of chemical modifications of the nucleotides
Sequence description – sequences of polymers, nucleic acids and proteins
Structural description – structure type
Citation – authors, title, journal, volume, pages, year

(*c*) *NMR structures*
Coordinate information – atomic coordinates for all models submitted
Experimental data for the deposition – spectrometer, types of NMR experiments, solution content, derivatization

(*d*) *X-ray structures*
Coordinate information – atomic (orthogonal or fractional) coordinates, occupancies and temperature factors for the asymmetric unit
Experimental data for the deposition – cell dimensions; space group for the X-ray structures, spectrometer, solution content *etc.*
Data collection description – radiation source and wavelength; data collection device; temperature; resolution range; total and unique number of reflections
Crystallization description – crystallization method; temperature; pH value; solution composition
Refinement information – program used; number of reflections used for refinement; data cutoff; resolution range; different *R* factors; refinement of temperature factors and occupancies

(*e*) *Derived information stored in the NDB*
Structure summary – descriptor briefly characterizing the structure
Distances – chemical bond lengths; virtual bonds between phosphorus atoms
Angles – valence angles, virtual angles involving phosphorus atoms
Torsion angles – backbone and side-chain torsion angles for nucleic acids and proteins; pseudo-rotational parameters for the sugar rings
Base pairing – detailed description of base pairing, classification by Leontis–Westhof (Leontis & Westhof, 2001) and Saenger (1984) with full identification of all pairs
Base morphology – base-morphology parameters calculated by *3DNA* (Alexandrov & Shindyalov, 2003) using the standard reference frame (Olson *et al.*, 2001)
Special geometrical features – nonbonded contacts, crystal-related geometries (for X-ray structures), symmetry-related coordinates, coordinates for symmetry-related strands, root-mean-square deviations from small-molecule standards for valence geometries
Sequence pattern statistics
Manually annotated structural features – type of helix, classification of double helix, presence of loops, bulges, three/four way junctions
Program-assigned structural features – presence of A/B/Z type of helix, presence of triple, quadruple helices, loops, bulges, three/four way junctions

dictionaries used for checking the structures were developed by the NDB Project from analyses of high-resolution, small-molecule structures (Clowney *et al.*, 1996; Gelbin *et al.*, 1996) from the Cambridge Structural Database (CSD; Allen *et al.*, 1979). The torsion-angle ranges for double-helical DNA forms were derived from an analysis of well resolved nucleic acid structures (Schneider *et al.*, 1997). One important outgrowth of these validation projects was the creation of the force constants and restraints that are now in common use for crystallographic refinement of nucleic acid structures (Parkinson *et al.*, 1996). In addition to geometry checks, the molecular model is checked against the experimental data using *SFCheck* (Vaguine *et al.*, 1999).

Once primary annotation and validation are complete, nucleic acid-specific structural and functional annotations are added. These vary from a broad characterization of nucleic acid conformation as double helical A, B or Z type, to a statement of the presence of a simple structural feature such as a bulge or a helical loop. Proteins from protein/nucleic acid complexes are annotated for their function in the complexes as structural proteins (including, *e.g.*, ribosomal and histone proteins), regulatory (including, *e.g.*, different types of transcription and translation co-factors) and enzymes (polymerases, topoisomerases, endonucleases *etc.*).

24.2.3. The database

24.2.3.1. Information content of the NDB

Structures available in the NDB include RNA and DNA oligonucleotides with two or more bases either alone or complexed with ligands, peptide nucleic acids, natural nucleic acids such as tRNA, and protein–nucleic acid complexes. The archive stores both primary and derived information about the structures (Table 24.2.3.1). The primary data include: the crystallographic coordinate data, structure factors, NMR constraints, and information about the experiments used to determine the structures, such as crystallization information, data collection and refinement statistics. Structural derived information, such as valence geometry, torsion angles and intermolecular contacts, is calculated and stored in the database. Structural and functional features annotated specifically for nucleic acid-containing structures are also loaded to the database.

Some structural features of nucleic acids have historically been derived by different algorithms, and it can be difficult to provide the most reliable values. Whenever possible, the NDB has tried to promote standards that allow structure comparison. An outstanding example of this difficulty was that different programs produced different values for base-morphology parameters (Lavery & Sklenar, 1989; Bansal *et al.*, 1995; Dickerson, 1998; Babcock *et al.*, 1994; Lu *et al.*, 1997). This meant that it was not possible to compare any two structures by using the numbers in the published literature and that it was necessary to recalculate these values for any analysis. To help resolve this problem, the NDB co-sponsored the *Tsukuba Workshop on Nucleic Acid Structure and Interactions* (Tsukuba, Japan, 1999). Key software developers in this field attended and resolved that a single reference frame would be used to calculate these values, and an agreement was reached about the definition of that reference frame (Olson *et al.*, 2001). This work fully quantitated the proposal for base morphology made previously at a meeting in Cambridge (Dickerson *et al.*, 1989). The NDB has recalculated these values for all the structures in the repository using the program *3DNA* (Lu & Olson, 2003), and they are available as output from NDB searches and prepared reports. All the programs have been amended so that they produce very similar values for the base-morphology parameters.

24.2.3.2. User web access

The core of the NDB project is a relational database that can be used to generate queries and reports about the stored structures. This key content of the NDB is supplemented by other services. An integrated web interface (http://ndbserver.rutgers.edu) provides wide and easy distribution of all NDB functionalities. The NDB home page links to the search and Atlas pages, coordinate files, standards for nucleic acid geometry, prepared reports about the database content, and pages with software tools for the calculation and analysis of nucleic acid structural properties.

24.2.3.3. Query capabilities

The NDB relational database has all the data organized into over 90 tables of summary, experimental and derived information, with each table containing five to 20 data items. For instance, the table *citation* contains all items relevant for related literature reference(s), *entity* contains critical items related to chemical entities observed in the structure and *entity_poly* contains the sequences of polymer(s) observed in the structure indexed to the entity table.

A web interface makes the query capabilities of the NDB as widely accessible as possible with a simple *NDB Search* and a more complex *Integrated Search*. The simple *NDB Search* provides a web form to query singular items, or combinations of items, such as NDB and PDB identification codes, author names and year of publication or release, experimental method, as well as chemical and structural characteristics such as the presence of chemical modification, the presence of non-Watson–Crick base pairing, the presence of different types of double helices (B/A/Z/right-handed), or the presence of a tetraplex or a three- or four-way junction. If no selection is made, the entire database is selected. After the query is executed, a list of structure IDs and descriptors that match the desired conditions is returned.

The *Integrated Search* provides the user with a wide range of options for structure selection and the use of Boolean operators to make complex queries. The interface offers pull-down menus to select key components of database entries: citation information, experiment type and details, molecule type (DNA, RNA, protein and ligand) present, presence of nucleic acid modifications, and nucleic acid structural conformation type (description, type and features).

Atlas pages (see below) and several different reports generated on the fly (Table 24.2.3.2) can be used to explore search results.

24.2.3.4. Atlas pages

A powerful reporting option is the NDB Atlas (Fig. 24.2.3.1). An Atlas report summarizes the most important characteristics of a structure with information about the authors, citation, sequence(s) of nucleic acids and proteins reported. Atlas entries contain an outline of the experimental information, a molecular view of the biologically relevant assembly of the reported structure, and links to tables of derived quantities such as torsion angles, base morphology and base pairing. Structural coordinates can be downloaded in PDB and mmCIF formats, and a link is provided to the PDB Structure Explorer page for more details.

Atlas pages are created for all entries directly from the NDB database and are organized by experimental method (X-ray and NMR), and by structure type into broad groups of structures (with possible overlaps) that allow effective browsing of the database. Atlas entries are also linked from database query results.

24.2.4. Distribution of information

In addition to querying and reporting capabilities, the NDB hosts coordinate and experimental data files, database reports, software programs, and other resources. They are available from the web (http://ndbserver.rutgers.edu/) as well as *via* the ftp server (ftp://ndbserver.rutgers.edu). An electronic help desk (ndbadmin@ndbserver.rutgers.edu) provides support.

Table 24.2.3.2
Reports automatically generated for searches of the NDB

Report name	Contains
Descriptor	NDB ID, structure description
NDB and PDB identifiers	NDB and PDB IDs
NDB status	NDB and PDB IDs, descriptor, authors, status, deposition and release dates
Citation information	Authors, title, journal, volume, pages, year
NA sequence	NDB ID, nucleic acid sequence, descriptor
Cell dimensions	Crystallographic cell constants, space group
Refinement data	Resolution limits and reflections used in refinement, *R* factor
Nucleic acid backbone torsion	Backbone torsions with unique residue identifier
Base-pair parameters	Buckle, propeller, opening, shear, stretch, stagger with unique pair identifier
Base-step parameters	Twist, rise, helical rise, tilt, roll, inclination, shift, slide, unique step identifier

24.2.4.1. Structural coordinates and experimental data

Structural coordinate files can be downloaded in mmCIF and PDB formats (Westbrook & Fitzgerald, 2009). Coordinates for the biological assembly of the molecule are provided for X-ray structures. When available, experimental files are made accessible to the community users.

24.2.4.2. Standards and software tools

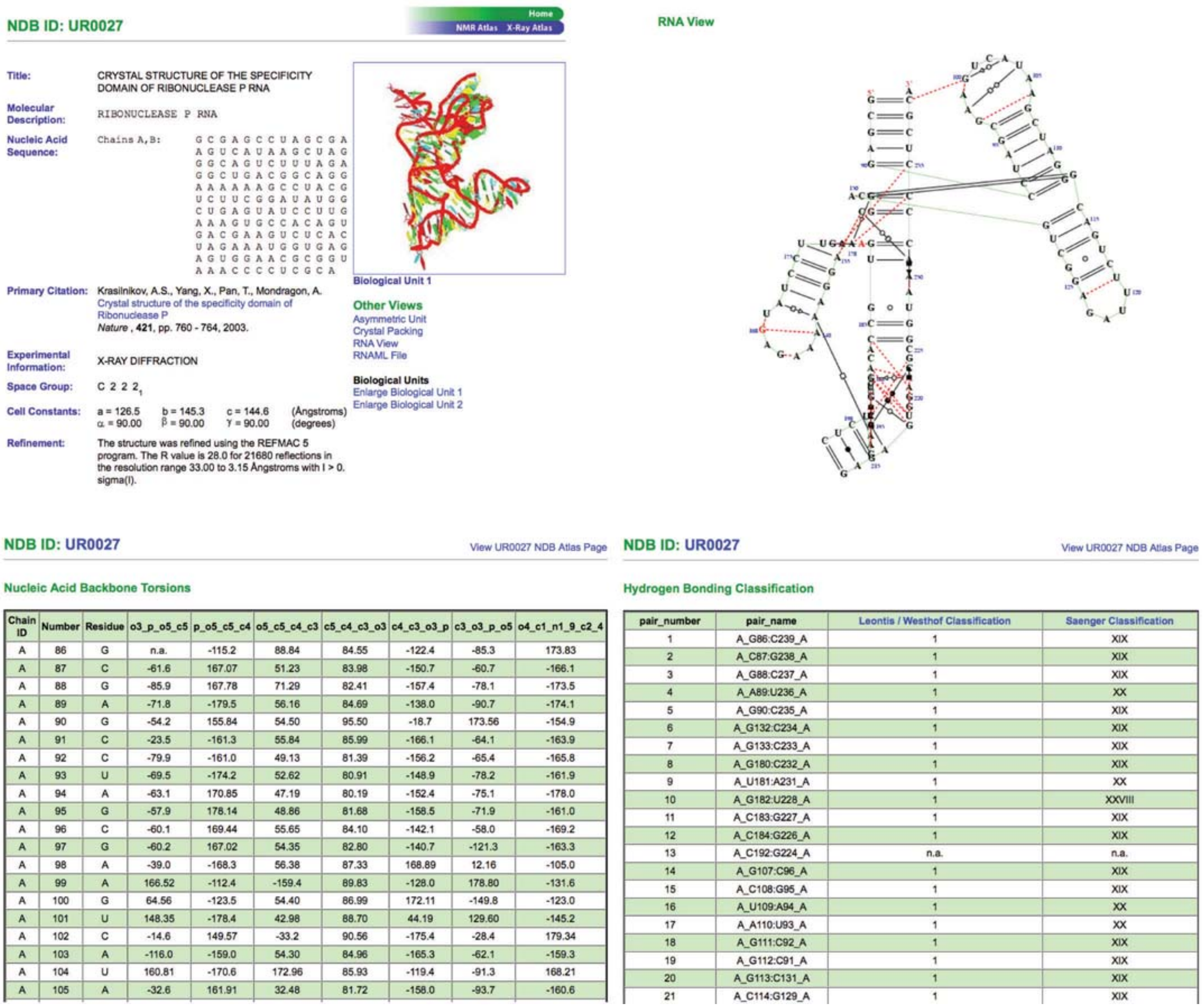
The *Standards* section contains dictionaries of standard (‘ideal’) geometries of nucleic acid components as well as their parameter files for *X-PLOR* (Brünger, 1992). It also provides a full explanation of the standard reference frame (Olson *et al.*, 2001).

The NDB links to several software *Tools*. With the recent increase in the number of RNA structures available, there have been attempts to establish systems for systematic analysis. The result of one of these studies has been the proposal of a classification scheme for the hydrogen bonds in the base pairs (Westhof & Fritsch, 2000; Leontis *et al.*, 2002). The syntax RNAML has also been proposed for representing RNA structural features (http://www-lbit.iro.umontreal.ca/rnaml/). The program *RNAViewer* (Yang *et al.*, 2003) uses Leontis–Westhof classification of base pairs and RNAML syntax to generate two-dimensional (2D) graphical representations of base pairing. *Base Pair Viewer* (*BPView*) visualizes selected base pairs or triplets. In both cases, the base pairs according to the Leontis–Westhof (Leontis & Westhof, 2001) classification are generated on the fly.

Two programs, *predictdnahth* (McLaughlin & Berman, 2003) and *HTHquery* (Ferrer-Costa *et al.*, 2005), analyse the protein DNA-binding helix-turn-helix (HTH) structural motif, and both can be used to determine whether a 3D protein target (from the archive or uploaded by the user) is a DNA-binding protein with the HTH motif.

Several software tools can be downloaded from the NDB site for local installation on the user's computer. *RNAViewer* (Yang *et al.*, 2003) and *S2S* (originally *RNAMLview*; Jossinet & Westhof, 2005) analyse base pairing in a PDB formatted file and draw its 2D representation. Three programs, *3DNA* (Lu & Olson, 2003), *Freehelix98* (Dickerson, 1998) and *hel_param* (Babcock *et al.*, 1993) are provided to compute base-pair parameters.

The NDB Project was funded by the National Science Foundation and the Department of Energy.



NDB ID: UR0027

Title: CRYSTAL STRUCTURE OF THE SPECIFICITY DOMAIN OF RIBONUCLEASE P RNA

Molecular Description: RIBONUCLEASE P RNA

Nucleic Acid Sequence: Chains A,B:

G C G A G C C U A G C G A
A G U C A U A A G C U A G
G G C A G U C U U U A G A
G G C U G A C G G C A G G
A A A A A A G C C U A C G
U C U U C G G A U A U G G
C U G A G U A U C C U U G
A A A G U G C C A C A G U
G A C G A A G U C U C A C
U A G A A A U G G U G A G
A G U G G A A C G C G G U
A A A C C C C U C G C A

Primary Citation: Krasilnikov, A.S., Yang, X., Pan, T., Mondragon, A.
Crystal structure of the specificity domain of Ribonuclease P
Nature , **421**, pp. 760 - 764, 2003.

Experimental Information: X-RAY DIFFRACTION

Space Group: C 2 2 2_1

Cell Constants: a = 126.5 b = 145.3 c = 144.6 (Ångstroms)
α = 90.00 β = 90.00 γ = 90.00 (degrees)

Refinement: The structure was refined using the REFMAC 5 program. The R value is 28.0 for 21680 reflections in the resolution range 33.00 to 3.15 Ångstroms with I > 0. sigma(I).

Biological Unit 1

Other Views
Asymmetric Unit
Crystal Packing
RNA View
RNAML File

Biological Units
Enlarge Biological Unit 1
Enlarge Biological Unit 2

NDB ID: UR0027 View UR0027 NDB Atlas Page

Nucleic Acid Backbone Torsions

Chain ID	Number	Residue	o3_p_o5_c5	p_o5_c5_c4	o5_c5_c4_c3	c5_c4_c3_o3	c4_c3_o3_p	c3_o3_p_o5	o4_c1_n1_9_c2_4
A	86	G	n.a.	-115.2	88.84	84.55	-122.4	-85.3	173.83
A	87	C	-61.6	167.07	51.23	83.98	-150.7	-60.7	-166.1
A	88	G	-85.9	167.78	71.29	82.41	-157.4	-78.1	-173.5
A	89	A	-71.8	-179.5	56.16	84.69	-138.0	-90.7	-174.1
A	90	G	-54.2	155.84	54.50	95.50	-18.7	173.56	-154.9
A	91	C	-23.5	-161.3	55.84	85.99	-166.1	-64.1	-163.9
A	92	C	-79.9	-161.0	49.13	81.39	-156.2	-65.4	-165.8
A	93	U	-69.5	-174.2	52.62	80.91	-148.9	-78.2	-161.9
A	94	A	-63.1	170.85	47.19	80.19	-152.4	-75.1	-178.0
A	95	G	-57.9	178.14	48.86	81.68	-158.5	-71.9	-161.0
A	96	C	-60.1	169.44	55.65	84.10	-142.1	-58.0	-169.2
A	97	G	-60.2	167.02	54.35	82.80	-140.7	-121.3	-163.3
A	98	A	-39.0	-168.3	56.38	87.33	168.89	12.16	-105.0
A	99	A	166.52	-112.4	-159.4	89.83	-128.0	178.80	-131.6
A	100	G	64.56	-123.5	54.40	86.99	172.11	-149.8	-123.0
A	101	U	148.35	-178.4	42.98	88.70	44.19	129.60	-145.2
A	102	C	-14.6	149.57	-33.2	90.56	-175.4	-28.4	179.34
A	103	A	-116.0	-159.0	54.30	84.96	-165.3	-62.1	-159.3
A	104	U	160.81	-170.6	172.96	85.93	-119.4	-91.3	168.21
A	105	A	-32.6	161.91	32.48	81.72	-158.0	-93.7	-160.6

NDB ID: UR0027 View UR0027 NDB Atlas Page

Hydrogen Bonding Classification

pair_number	pair_name	Leontis / Westhof Classification	Saenger Classification
1	A_G86:C239_A	1	XIX
2	A_C87:G238_A	1	XIX
3	A_G88:C237_A	1	XIX
4	A_A89:U236_A	1	XX
5	A_G90:C235_A	1	XIX
6	A_G132:C234_A	1	XIX
7	A_G133:C233_A	1	XIX
8	A_G180:C232_A	1	XIX
9	A_U181:A231_A	1	XX
10	A_G182:U228_A	1	XXVIII
11	A_C183:G227_A	1	XIX
12	A_C184:G226_A	1	XIX
13	A_C192:G224_A	n.a.	n.a.
14	A_G107:C96_A	1	XIX
15	A_C108:G95_A	1	XIX
16	A_U109:A94_A	1	XX
17	A_A110:U93_A	1	XX
18	A_G111:C92_A	1	XIX
19	A_G112:C91_A	1	XIX
20	A_G113:C131_A	1	XIX
21	A_C114:G129_A	1	XIX

Figure 24.2.3.1
NDB Atlas entry for a structure of ribonuclease P RNA, UR0027 (Krasilnikov *et al.*, 2003). Clockwise from the upper left: the summary Atlas page, schematic 2D drawing of base pairing as calculated and drawn by *RNAView* (Yang *et al.*, 2003), hydrogen-bonding classification, and torsion angles.

This chapter is an updated version of Schneider *et al.* (2009). Reprinted with permission of John Wiley & Sons, Inc.

References

Alexandrov, N. & Shindyalov, I. (2003). *PDP: protein domain parser*. *Bioinformatics*, **19**, 429–430.

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information*. *Acta Cryst.* B**35**, 2331–2339.

Babcock, M. S., Pednault, E. P. & Olson, W. K. (1993). *Nucleic acid structure analysis: a users guide to a collection of new analysis programs*. *J. Biomol. Struct. Dyn.* **11**, 597–628.

Babcock, M. S., Pednault, E. P. & Olson, W. K. (1994). *Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures*. *J. Mol. Biol.* **237**, 125–156.

Bansal, M., Bhattacharyya, D. & Ravi, B. (1995). *Nuparm and nucgen: software for analysis and generation of sequence dependent nucleic acid structures*. *CABIOS*, **11**, 281–287.

Berman, H., Henrick, K. & Nakamura, H. (2003). *Announcing the worldwide Protein Data Bank*. *Nat. Struct. Biol.* **10**, 980.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). *The Nucleic Acid Database. A comprehensive relational database of three-dimensional structures of nucleic acids*. *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *The Protein Data Bank: a computer-based archival file for macromolecular structures*. *J. Mol. Biol.* **112**, 535–542.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D. & Fitzgerald, P. M. (1997). *Macromolecular Crystallographic Information File*. *Methods Enzymol.* **277**, 571–590.

Brünger, A. T. (1992). *X-plor. A System for X-ray Crystallography and NMR*. Version 3.1. Yale University, USA.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: nitrogenous bases*. *J. Am. Chem. Soc.* **118**, 509–518.

Dickerson, R. E. (1998). *DNA bending: the prevalence of kinkiness and the virtues of normality*. *Nucleic Acids Res.* **26**, 1906–1926.

Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., von Kitzing, E., Lavery, R., Nelson, H. C. M., Olson, W., Saenger, W., Shakked, Z., Sklenar, H., Soumpasis, D. M., Tung, C.-S., Wang, A. H.-J. & Zhurkin, V. B. (1989). *Definitions and nomenclature of nucleic acid structure parameters*. *EMBO J.* **8**, 1–4.

Ferrer-Costa, C., Shanahan, H. P., Jones, S. & Thornton, J. M. (2005). *HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif*. *Bioinformatics*, **21**, 3679–3680.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005*a*). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, pp. 295–443. Dordrecht: Springer.

Fitzgerald, P. M. D., Westbrook, J. D., Bourne, P. E., McMahon, B., Watenpaugh, K. D. & Berman, H. M. (2005*b*). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, pp. 144–198. Dordrecht: Springer.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *Geometric parameters in nucleic acids: sugar and phosphate constituents*. *J. Am. Chem. Soc.* **118**, 519–528.

Jossinet, F. & Westhof, E. (2005). *Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure*. *Bioinformatics*, **21**, 3320–3321.

Krasilnikov, A. S., Yang, X., Pan, T. & Mondragón, A. (2003). *Crystal structure of the specificity domain of ribonuclease P*. *Nature (London)*, **421**, 760–764.

Lavery, R. & Sklenar, H. (1989). *Defining the structure of irregular nucleic acids: conventions and principles*. *J. Biomol. Struct. Dyn.* **6**, 655–667.

Leontis, N. B., Stombaugh, J. & Westhof, E. (2002). *The non-Watson–Crick base pairs and their associated isostericity matrices*. *Nucleic Acids Res.* **30**, 3497–3531.

Leontis, N. B. & Westhof, E. (2001). *Geometric nomenclature and classification of RNA base pairs*. *RNA*, **7**, 499–512.

Lu, X. J., El Hassan, M. A. & Hunter, C. A. (1997). *Structure and conformation of helical nucleic acids: analysis program (SCHNAaP)*. *J. Mol. Biol.* **273**, 668–680.

Lu, X. J. & Olson, W. K. (2003). *3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures*. *Nucleic Acids Res.* **31**, 5108–5121.

McLaughlin, W. A. & Berman, H. M. (2003). *Statistical models for discerning protein structures containing the DNA-binding helix-turn-helix motif*. *J. Mol. Biol.* **330**, 43–55.

Moore, P. B. (2001). *The ribosome at atomic resolution*. *Biochemistry*, **40**, 3243–3250.

Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C., Heinemann, U., Lu, X.-J., Neidle, S., Shakked, Z., Sklenar, H., Suzuki, M., Tung, C.-S., Westhof, E., Wolberger, C. & Berman, H. M. (2001). *A standard reference frame for the description of nucleic acid base-pair geometry*. *J. Mol. Biol.* **313**, 229–237.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *New parameters for the refinement of nucleic acid-containing structures*. *Acta Cryst.* D**52**, 57–64.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*. New York: Springer-Verlag.

Schneider, B., de la Cruz, J., Feng, Z., Chen, L., Westbrook, J., Yang, H., Young, J., Zardecki, C. & Berman, H. M. (2009). *Structural Bioinformatics*, 2nd ed., edited by J. Gu & P. E. Bourne, pp. 293–303. Hoboken: John Wiley & Sons, Inc.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Conformations of the sugar–phosphate backbone in helical DNA crystal structures*. *Biopolymers*, **42**, 113–124.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model*. *Acta Cryst.* D**55**, 191–205.

Westbrook, J., Feng, Z., Burkhardt, K. & Berman, H. M. (2003). *The Protein Data Bank*. *Nucleic Acids Res.* **28**, 235–242.

Westbrook, J., Yang, H., Feng, Z. & Berman, H. M. (2005). *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall & B. McMahon, pp. 539–543. Dordrecht: Springer.

Westbrook, J. D. & Bourne, P. E. (2000). *STAR/mmCIF: an ontology for macromolecular structure*. *Bioinformatics*, **16**, 159–168.

Westbrook, J. D. & Fitzgerald, P. M. D. (2009). *Structural Bioinformatics*, 2nd ed., edited by P. E. Bourne & J. Gu, pp. 271–291. Hoboken: John Wiley & Sons, Inc.

Westhof, E. & Fritsch, V. (2000). *RNA folding: beyond Watson–Crick pairs*. *Structure*, **8**, R55–R65.

Yang, H., Jossinet, F., Leontis, N., Chen, L., Westbrook, J., Berman, H. M. & Westhof, E. (2003). *Tools for the automatic identification and classification of RNA base pairs*. *Nucleic Acids Res.* **31**, 3450–3460.

International Tables for Crystallography (2012). Vol. F, Chapter 24.3, pp. 838–842.

Chapter 24.3. The Biological Macromolecule Crystallization Database

D. T. Gallagher and M. Tung

24.3.1. Introduction

The crystallization of a biological macromolecule is a key step in determining its three-dimensional structure by X-ray diffraction. Even for proteins of known structure, it is very difficult to make predictions about crystal-growth conditions or crystal properties. Hence, in crystallizing macromolecules, empirical procedures are used that take advantage of the knowledge gained from past successes. Usually a large set of trials is carried out that varies parameters such as pH, temperature, ionic strength and macromolecule concentration. The number of experiments required for success is variable. In many cases the search ends quickly, either because the right choices were made early or because crystallization occurs over a broad range of conditions. Unfortunately, in other cases, a large number of experiments are required before the discovery of crystallization conditions, and in some cases no crystal conditions are found even after exhaustive searching.

After more than 50 years of experience in the production of diffraction-quality crystals, there is still no generally accepted strategy for searching for the crystal-growth conditions for a biological macromolecule. However, a number of systematic procedures and strategy suggestions have been put forth (*e.g.* McPherson, 1976; Blundell & Johnson, 1976; Carter & Carter, 1979; McPherson, 1982; Gilliland & Davies, 1984; Gilliland *et al.*, 1994, 1996; McPherson, 1999). Most current strategies employ a version of the ‘fast screen’ first popularized by Jancarik & Kim (1991). Fast screens are sets of preformulated solutions similar to those that have frequently produced crystals of other proteins in the past. Crystals are often found quickly in such experiments, but failure results in the need for a more general approach.

The purpose of the Biological Macromolecule Crystallization Database (BMCD) is to provide comprehensive information to facilitate the development of crystallization strategies for the production of crystals suitable for X-ray structural investigations (Gilliland & Davies, 1984). The BMCD includes entries for all classes of biological macromolecules for which diffraction-quality crystals have been obtained, including proteins, nucleic acids and their various complexes.

24.3.2. History of the BMCD

The BMCD has its roots in work that was initiated in Dr David Davies’ laboratory at NIH in the late 1970s and early 1980s (Gilliland & Davies, 1984). While working on a variety of frustrating protein-crystallization problems, a large body of crystallization information was extracted from the literature. This eventually led to a systematic search of the literature and a compilation of data that included almost all of the crystallization reports of biological macromolecules available at the time. In 1983 the data, as an ASCII file, were submitted to the Protein Data Bank (PDB; see Chapter 24.1) for public distribution. The data included the crystallization conditions for 1025 crystal forms of more than 616 biological macromolecules.

In 1987, with assistance from the National Institute of Standards and Technology (NIST) Standard Reference Data Program, the data were incorporated into a searchable database and distributed with software that made it accessible using a personal computer. The database was released to the public in 1989 as the NIST/CARB (Center for Advanced Research in Biotechnology) Biological Macromolecule Crystallization Database, version 1.0 (Gilliland, 1988). In 1990, a second version was released (Gilliland & Bickham, 1990), and in 1994 the BMCD began including data from microgravity crystal-growth studies carried out in orbit (Gilliland *et al.*, 1994). Soon afterwards, the BMCD was migrated to a UNIX platform to facilitate Internet access, and the software was rewritten to enable direct import of data from the PDB. During the period 1997–1998 this reprogramming superceded new data acquisition. Hence, there is a local minimum of crystal entries for that period. In 1999, data acquisition resumed and the size of the BMCD began to increase sharply. In 2008, the BMCD released version 4.01 (Tung & Gallagher, 2009), using the open-source database server PostgreSQL 8.1.3. This version also includes new search features capable of searching for arbitrary text and for ranges of five numeric parameters (pH *etc.*) as described below.

24.3.3. BMCD data

The BMCD stores information in 40 different data types generally divisible into three groups relating to the macromolecule, the crystallization and the crystal itself. Under macromolecule are the name, aliases, biological source genus, tissue *etc.*, mutations, and numbers and sizes of subunit types. The macromolecule sequence is also included but is not yet searchable. The crystal-growth information contains the temperature, the pH, the concentrations of all chemical components including the macromolecule, growth time and the method. Crystal data include the space group, unit-cell parameters, molecules per cell and diffraction resolution. Also included are the V_m and solvent-fraction values. An additional section for each entry records references to published literature.

The BMCD4 (current release, version 4) contains two classes of entries. Those that belonged to BMCD version 3 (about 3500 entries, generally corresponding to information added before 1996) were obtained manually and tend to have complete information relating to crystal growth, including method-specific details. Approximately half of these entries derive from literature reports only and there is no structure in the PDB that directly corresponds to them. These entries have BMCD ID codes that begin with the letter M. The second class of entries are those that are new in version 4, and were obtained by retrieving and parsing data from the PDB roughly covering the period 1997–2007. This represents the new model of data acquisition, utilizing the PDB’s RSYNC download utility to obtain XML files for each entry. The XML files are then processed by custom Java scripts to select data items of interest and convert them into database tables. This conversion involves extensive use of text-parsing scripts and human attention, as described for a similar data-acquisition project by Peat *et al.* (2005). For this processing, filtering and error checking are required in order to correctly interpret unformatted PDB information, especially the crystal conditions.

Copyright © 2012 International Union of Crystallography

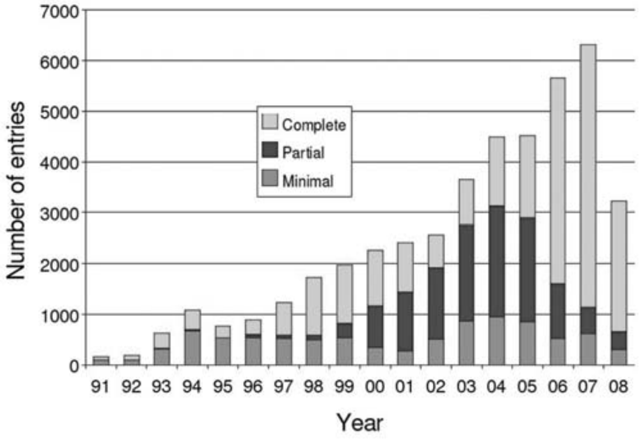


Figure 24.3.3.1
Trends in the completeness of BMCD crystallization conditions.

The resulting entries correspond directly to PDB structures, and the first four characters of the BMCD ID code are the same as the corresponding PDB code. Because the PDB deposition does not require crystal-growth information, many BMCD entries imported from the PDB contain incomplete information. BMCD entries can be divided into three groups according to the completeness of their crystal-growth information: those with 'complete' information, those listing chemicals but not concentrations ('partial'), and those that lack even chemical names but still have some information, such as pH ('minimal'). For a set of 43 698 BMCD entries including the currently available version 4.01 as well as 29 326 entries currently undergoing validation processing, the distribution of these three groups is shown in Fig. 24.3.3.1. As can be seen, most of the entries from 2004 and 2005 are incomplete, but more recently the trend appears to be in favour of recording complete conditions.

An additional complication arises from the wide assortment of synonyms and misspellings in the raw data. Since statistical analysis of chemical conditions requires standardization of chemical names, we are developing methods to interpret and convert both synonyms (*e.g.*, AmSO4, A.S. and AmS for ammonium sulfate) and common misspellings into standard chemical names. This process results in a large reduction in the number of unique chemical names, dominated by a few chemicals that have many synonyms and many misspellings. Efforts to standardize the input and archiving of crystallization information are also underway at major crystallography journals (Einspahr & Guss, 2005).

24.3.4. Web interface

The BMCD is an internet-accessible resource available through the website http://xpdb.nist.gov:8060/BMCD4. The web interface provides a convenient mechanism for browsing through the data contained in the BMCD. The user can examine individual entries, lists of entries selected by customizable search criteria or statistical trends based on the numbers of entries for various properties and parameter ranges. In addition, specific references are listed for most entries.

Two types of search are possible. A simple text search is initiated by input on the home page. For example, a search for 'DTT OR mercaptoethanol' (single quotes are optional in the actual input; the search string is case-insensitive) returns about a thousand entries that contain one or the other (or both) of these reducing agents, while a search for 'dtt mercaptoethanol' has the same result (OR is implicit). A search for "double mutant" returns the 13 entries that contain this phrase, while the search 'double AND mutant' returns a slightly larger superset that includes any entry with both the search terms. The asterisk character functions as a wildcard, enabling the search for "tetra*" to yield a large set that includes tetragonal crystals as well as explicitly tetrameric proteins. Searching for 'mono* NOT monoclinic' retrieves non-monoclinic entries that contain the words monomeric, mononucleotide *etc*. The query syntax is explained on linked information pages.

An advanced search is offered from its own query page, which accepts input in the form of numeric ranges for any of the five parameters: macromolecule concentration, pH, temperature, resolution and year of publication. BMCD entries with parameters that lie within all the specified ranges are then listed as output. In addition, the advanced search accepts text input (like the simple search) to further specify the target set. For example, using the advanced search, one could identify the 14 entries that contain the text string 'adenosine' and have pH between 3 and 6.6, and also have a diffraction resolution between 0.1 and 1.85 Å.

One additional search feature is the ability to request text matches within specific fields. Each entry has distinct text fields for protein name, organism name, space group, chemical names *etc.* (the full list of searchable fields, with examples, is linked to the search page), and these can be searched independently using a colon syntax. For example, the query

```
title: activation
```

will find any entry whose publication title includes the word 'activation'. The query

```
title:″activation domain″
```

will return the smaller set where the title includes this phrase. Multiple field searches may be combined using Boolean operators. A term or phrase not preceded by a field name will be searched through the entire entry (general search). However, field searching and non-field searching cannot be combined in the same query. The way to combine field and general searching is to use the "Content:" field, which is effectively a general search over all fields. Field names are case-specific; all must be completely lower-case, except "Content".

Here are a few examples of correct syntax for field searches:

```
title:″HIV-1 protease″ AND spgrp:P61
title:antibody AND common_name:mouse
title:antibody AND Content:mouse AND NOT
    common_name:mouse
chem_name:aden * AND (author:mckay OR author:steitz)
```

The output of a search begins with the number of entries found, followed by a linked list of their BMCD codes, molecular names and biological sources. The molecular-name field usually includes the scientific name of the molecule along with common synonyms as previously described (Gilliland *et al.*, 1994). By clicking on the ID code of one of the entries all the data pertaining to that entry are displayed.

24.3.5. Reproducing published crystallization procedures

The BMCD contains the information required to reproduce previously reported crystals for a biological macromolecule. Usually, a later batch of macromolecule will not behave identically to previous samples and some optimizing is required, beginning with the reported crystallization conditions. The conditions may be simple to reproduce, but differences in the isolation and purification procedures, reagents, and crystallization methodology of different laboratories can dramatically influence the results. The crystallization conditions in the BMCD should be considered as a good starting point for the search or optimization that may require experiments that vary pH, macromolecule and reagent concentrations, and temperature.

An attempt to reproduce crystals of an isozyme of glutathione *S*-transferase from rat liver (Sesay *et al.*, 1987) is used to illustrate these points. The original crystallization conditions, for the enzyme purified from liver tissue, were archived as entry M0P3. Several years later the same enzyme was cloned and expressed in *Escherichia coli* for further structural studies. The crystals of the original enzyme grew in 3 to 5 days from vapour-diffusion experiments at 4 °C, with droplets containing a protein concentration of 11 mg ml^{-1}, 0.46% β-octylglucoside, 30–37% saturated ammonium sulfate and 0.1 M phosphate buffer pH 6.9 equilibrated against well solutions containing 60–74% ammonium sulfate.

The recombinant enzyme required an optimization of these conditions to produce large single crystals. The recombinant protein crystallized best at the same temperature, with droplets containing a protein concentration of 12 mg ml^{-1}, 0.2% β-octylglucoside, 20–25% saturated ammonium sulfate, 1 mM EDTA and 0.025 M TrisHCl, pH 8.0 equilibrated against well solutions containing 40–50% ammonium sulfate. Both crystallization protocols required the presence of 1 mM (9*R*,10*R*)-9-*S*-glutathionyl-10-hydroxy-9,10-dihydrophenanthrene, a product inhibitor. Thus, in this example, most of the chemical concentrations shifted, and the pH shifted so much that a different buffer was used. Optimized crystals of the recombinant enzyme grew in 5 to 10 days (Ji *et al.*, 1994).

24.3.6. Crystallization screens

A crystallization screen is a set of formulations designed to be mixed with macromolecule solutions as an efficient search for crystallogenic conditions; several are now available commerically as sets of premixed solutions (*e.g.* Cudney *et al.*, 1994). Since the introduction of the fast screen by Jancarik & Kim (1991), almost all attempts to crystallize a new protein begin with a screen of one form or another. Early screens was based on the ideas put forth by Carter & Carter (1979) in their discussion of the use of incomplete factorial experiments to limit the search. The first screens were quite general and applicable to a wide range of biological macromolecules, but screens based on specific classes of molecules such as RNA (Scott *et al.*, 1995) or protein complexes (Radaev *et al.*, 2006) soon developed.

The design of crystal screens is based on previously successful crystallizations. The BMCD has been used extensively for the development of general screens, and can be used to facilitate the development of screens for specific classes of macromolecules. For example, if it were desired to produce a screen for endonucleases, a search of the BMCD would reveal the ranges of key parameters and the most prevalent reagents that have been used to crystallize these enzymes. From an examination of these parameters, a subset of potential crystallization conditions comprising an endonuclease screen could be developed. Another type of focused screen could be one designed to feature a particular crystallant. For example, if L-ornithine were theorized to promote the crystallization of proteins, a screen featuring this additive could be designed by surveying the BMCD for relevant precedents and parameter ranges. Similar methods could be used to focus the initial search for any specific crystallization project.

24.3.7. A general crystallization procedure

The information in the BMCD has been incorporated into general procedures for the crystallization of biological macromolecules not previously crystallized (Gilliland, 1988; Gilliland & Bickham, 1990; Gilliland *et al.*, 1994, 1996). One such general procedure is shown in Fig. 24.3.7.1. Briefly, in this procedure the purified biological macromolecule is concentrated (if possible) to 10 to 25 mg ml^{-1} and dialysed into 10 to 25 mM buffer at a neutral pH or at a pH favouring the solubility of the biopolymer. Other stabilizing agents such as EDTA and/or dithiothreitol may be included at low concentrations to stabilize the biological macromolecule during the crystallization trials. If the biomolecule requires some salt for stability or solubility, then it should be included, but ideally the additional components in the protein solution are kept to a minimum as they limit the search space.

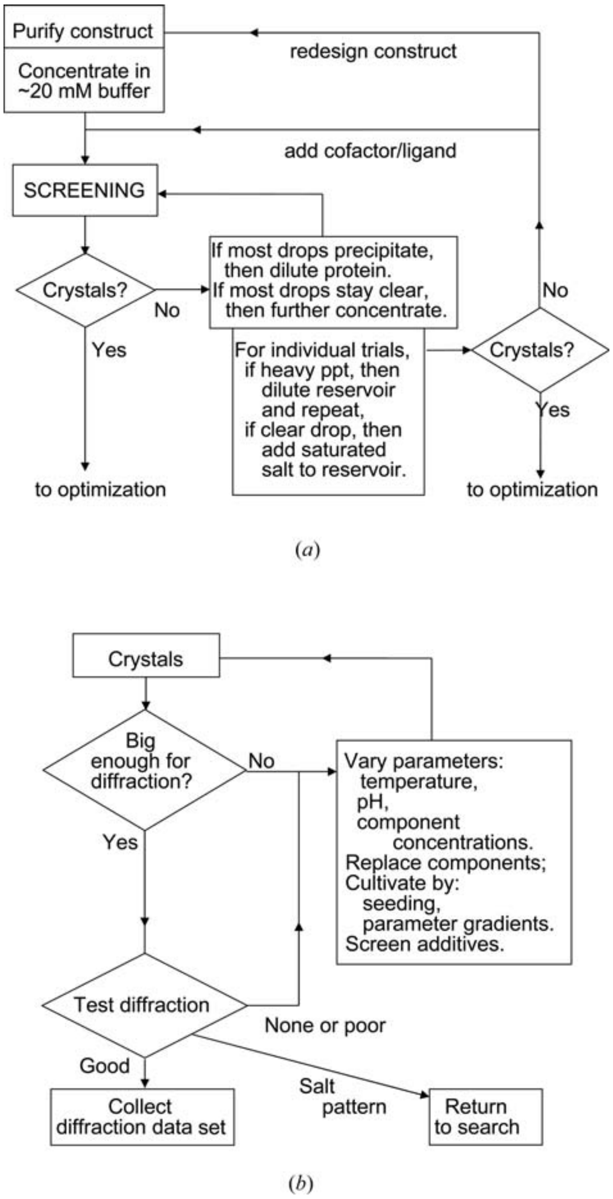

Purify construct
Concentrate in ~20 mM buffer
redesign construct
add cofactor/ligand
SCREENING
Crystals?
No
Yes
to optimization
If most drops precipitate, then dilute protein.
If most drops stay clear, then further concentrate.
For individual trials, if heavy ppt, then dilute reservoir and repeat, if clear drop, then add saturated salt to reservoir.
Crystals?
No
Yes
to optimization
(a)
Crystals
Big enough for diffraction?
No
Yes
Vary parameters: temperature, pH, component concentrations. Replace components; Cultivate by: seeding, parameter gradients. Screen additives.
Test diffraction
None or poor
Good
Salt pattern
Collect diffraction data set
Return to search
(b)

Figure 24.3.7.1
A general crystallization strategy based on the data contained in the BMCD. The overall strategy comprises a search phase (*a*) and an optimization phase (*b*).

Once the macromolecule solution has been prepared, commercial or customized fast screens are carried out using vapour-diffusion experiments. If no crystals appear, it is important to recognize that some trials may contain favourable chemical conditions but fail to produce crystals simply because the concentrations are too high or too low. Thus, each screen trial should be considered not as a single point in composition space but as a vector to be scanned, and any given trial cannot be considered to have failed until its region of transition (between clear drop and precipitation) is observed. If most of the trials produce heavy precipitate, then systematic dilution (either of the reservoirs or of the protein) is advised. If on the other hand most of the drops remain clear, then either more concentrated protein or more dehydrating reservoir conditions will help to bring the protein to its solubility limit in that milieu, and hence to its potential crystal-nucleation point. Individual trials can be manipulated to ensure that no potential crystallization conditions are missed simply because of too high or too low concentation.

Often the first crystals to appear are small, compound or otherwise poor in quality. In this case the optimization phase (Fig. 24.3.7.1*b*) begins by attempting to simplify the crystal conditions (can any components be omitted?), and then systematically varying the crystallization parameters (pH, temperature, chemical concentrations) in the hope of improving size and quality. These experiments generally incorporate controls corresponding to the best crystals obtained so far, so that reproducibility is continually assessed. When crystals are large enough (usually about 0.02 mm), diffraction is tested. Beyond optimizing existing parameters, new components can be substituted (*e.g.*, replacing PEG 4K with PEG 5K MME) or introduced *de novo*, using crystallizations of similar proteins or crystallizations that utilized similar conditions in the BMCD as a guide. Microseeding (often useful to obviate nucleation and to control populations) or macroseeding (to cultivate large, high-quality single crystals) may also be required to optimize crystal growth (McPherson, 1982, 1999).

If the fast screens produce no crystals, a more thorough approach can be undertaken. An analysis of the BMCD data reveals that out of the large number of reagents used as precipitating agents, a small set accounts for the majority of the crystals observed. The pH range for all crystals is quite large, but most proteins crystallize between pH 4.0 and 9.0. Even though temperature can be an important factor, crystallization experiments are usually set up at room (20 °C) or cold-room (5 °C) temperatures. Protein concentration varies quite markedly, but it appears that most experiments use from 5 to 15 mg ml^{-1}. After examining the data in the BMCD, the precipitating agents ammonium sulfate, polyethylene glycol 4000, 2-methyl-2,4-pentanediol and sodium–potassium phosphate might be selected for custom-screening efforts, with initial trials restricted to a pH range of 4.0 to 9.0 and temperatures of 5 and 20 °C. As a prescreening solubility assay, a small amount (2 µl) of the protein is mixed with several concentrations of each of the selected precipitants, and buffered at pH 4.5, 6.0 and 8.0, at both temperatures, monitoring by microscope for precipitation. This establishes the concentration ranges for the reagents for setting up hanging-drop (or any other commonly used technique) experiments. Next, separate sets of experiments that would sample the pH range in steps of 0.5 and reagent concentrations near, at and above those that induce precipitation of the protein would be set up at temperatures of 5 and 20 °C. The assessment of the results of experiments after periodic observations may show (for example by an abrupt precipitation at a particular reagent concentration, pH and/or temperature) a need for finer sampling of any or all of the parameters near the observed discontinuity. In parallel, or if the crystallization trials just described are unsuccessful, another set of experiments can be carried out that include the addition of small quantities of ligands, products, substrate, substrate analogues, monovalent or divalent cations, organic reagents *etc.* to the crystallization mixtures. If this does not prove fruitful, additional reagents may be selected with the aid of the BMCD and new experiments initiated.

24.3.8. The future of the BMCD

Further developments for the BMCD are in progress, with the general goal of increasing its capacity for data analysis, thereby facilitating both scientific understanding and practical applications. It is anticipated that text-parsing and error-correction scripts will soon enable efficient regular import of data from the PDB. This will enable the automatic parsing of all crystallization information into specific chemicals with numerically stored range-searchable concentrations to facilitate detailed statistical analysis. Additionally, future revisions of the database will incorporate taxonomic information on source organisms, classification of proteins, sequence analysis and more powerful searches with more user control of search outputs. The capabilities of the web resource will be expanded to include tools for the development of strategies for new crystallization problems.

The database and its publication are supported by the US National Institute of Standards and Technology. Assistance with database development and support from Gary Gilliland, Jane Ladner and Robert Goldberg is gratefully acknowledged. Certain commercial equipment, instruments and materials are identified in this paper in order to specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials and equipment identified are necessarily the best available for the purpose.

References

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. New York: Academic Press.

Carter, C. W. Jr & Carter, C. W. (1979). *Protein crystallization using incomplete factorial experiments*. *J. Biol. Chem.* **254**, 12219–12223.

Cudney, R., Patel, S., Weisgraber, K., Newhouse, Y. & McPherson, A. (1994). *Screening and optimization strategies for macromolecular crystal growth*. *Acta Cryst.* D**50**, 414–423.

Einspahr, H. & Guss, M. (2005). *Editorial*. *Acta Cryst.* F**61**, 1–2.

Gilliland, G. L. (1988). *A biological macromolecule crystallization database: a basis for a crystallization strategy*. *J. Cryst. Growth*, **90**, 51–59.

Gilliland, G. L. & Bickham, D. (1990). *The Biological Macromolecule Crystallization Database: a tool to assist the development of crystallization strategies*. *Methods*, **1**, 6–11.

Gilliland, G. L. & Davies, D. R. (1984). *Protein crystallization: the growth of large-scale single crystals*. *Methods Enzymol.* **104**, 370–381.

Gilliland, G. L., Tung, M., Blakeslee, D. M. & Ladner, J. E. (1994). *Biological Macromolecule Crystallization Database, Version 3.0: new features, data and the NASA archive for protein crystal growth data*. *Acta Cryst.* D**50**, 408–413.

Gilliland, G. L., Tung, M. & Ladner, J. (1996). *The Biological Macromolecule Crystallization Database and NASA Protein Crystal Growth Archive*. *J. Res. Natl Inst. Stand. Technol.* **101**, 309–320.

Jancarik, J. & Kim, S.-H. (1991). *Sparse matrix sampling: a screening method for crystallization of proteins*. *J. Appl. Cryst.* **24**, 409–411.

Ji, X., Johnson, W. W., Sesay, M. A., Dickert, L., Prasad, S. M., Ammon, H. L., Armstrong, R. N. & Gilliland, G. L. (1994). *Structure of the xenobiotic substrate binding site of a glutathione S-transferase as* *revealed by X-ray crystallographic analysis of product complexes with the diastereomers of 9-(S-glutathionyl)-10-hydroxy-9,10-dihydrophenanthrene*. *Biochemistry*, **33**, 1043–1052.

McPherson, A. (1976). *The growth and preliminary investigation of protein and nucleic acid crystals for X-ray diffraction analysis*. *Methods Biochem. Anal.* **23**, 249–345.

McPherson, A. (1982). *Preparation and Analysis of Protein Crystals*. New York: Wiley.

McPherson, A. (1999). *Crystallization of Biological Macromolecules*. New York: Cold Spring Harbor Laboratory Press.

Peat, T. S., Christopher, J. A. & Newman, J. (2005). *Tapping the Protein Data Bank for crystallization information*. *Acta Cryst.* D**61**, 1662–1669.

Radaev, S., Li, S. & Sun, P. D. (2006). *A survey of protein–protein complex crystallizations*. *Acta Cryst.* D**62**, 605–612.

Scott, W. G., Finch, J. T., Grenfell, R., Fogg, J., Smith, T., Gait, M. J. & Klug, A. (1995). *Rapid crystallization of chemically synthesized hammerhead RNAs using a double screening procedure*. *J. Mol. Biol.* **250**, 327–332.

Sesay, M. A., Ammon, H. L. & Armstrong, R. N. (1987). *Crystallization and a preliminary X-ray diffraction study of isozyme 3–3 of glutathione S-transferase from rat liver*. *J. Mol. Biol.* **197**, 377–378.

Tung, M. & Gallagher, D. T. (2009). *The Biomolecular Crystallization Database Version 4: expanded content and new features*. *Acta Cryst.* D**65**, 18–23.

International Tables for Crystallography (2012). Vol. F, Chapter 25.1, pp. 845–872.

PART 25. A HISTORICAL PERSPECTIVE

Chapter 25.1. How the structure of lysozyme was actually determined

C. C. F. Blake, R. H. Fenn, L. N. Johnson, D. F. Koenig, G. A. Mair, A. C. T. North, J. W. H. Oldham, D. C. Phillips, R. J. Poljak, V. R. Sarma and C. A. Vernon

25.1.1. Introduction

For protein crystallographers, the year 1960 was the spring of hope. The determination of the three-dimensional structure of sperm-whale myoglobin at 2 Å resolution (Kendrew *et al.*, 1960) had shown that such analyses were possible, and the parallel study of horse haemoglobin at 5.5 Å resolution (Perutz *et al.*, 1960) had shown that even low-resolution studies could, under favourable circumstances, reveal important biological information. All seemed set for a dramatic expansion in protein studies.

At the Royal Institution in London, two of us (CCFB and DCP) had used the laboratory-prototype linear diffractometer (Arndt & Phillips, 1961) to extend the myoglobin measurements to 1.4 Å resolution for use in refinement of the structure (Watson *et al.*, 1963), and we had begun a detailed study of irradiation damage in the myoglobin crystals (Blake & Phillips, 1962). Meanwhile, David Green, an early contributor to the haemoglobin work (Green *et al.*, 1954), and ACTN had initiated a study of β-lactoglobulin (Green *et al.*, 1956) and worked together on oxyhaemoglobin before Green went to the Massachusetts Institute of Technology (MIT) in 1959 on leave for a year. At roughly this time, many of the participants in the myoglobin and haemoglobin work at Cambridge went off to other laboratories to initiate or reinforce other studies. Thus, Dick Dickerson went with Larry Steinrauf to the University of Illinois, Urbana, to start a study of the triclinic crystals of hen egg-white lysozyme.

RJP went to MIT from the Argentine as a post-doctoral fellow in 1958 and worked initially with Martin Buerger. In 1959 he transferred to Alex Rich's laboratory and there he soon came into contact with a number of veterans of the myoglobin and haemoglobin work. In addition to David Green were Howard Dintzis, who had discovered a number of the important heavy-atom derivatives of myoglobin (Bluhm *et al.*, 1958) and was now on the staff at MIT, and David Blow, who had first used multiple isomorphous replacement and anomalous scattering to determine haemoglobin phases (Blow, 1958) and was on leave from Cambridge. The influence of these people, combined with lectures by John Kendrew and then by Max Perutz on visits to MIT, soon convinced RJP that working on the three-dimensional structures of proteins was the most challenging and fruitful research that a crystallographer could undertake. Dintzis, in particular, persuaded him that preparing heavy-atom derivatives was no great problem, and Blow urged him to look for commercially available proteins that were known to crystallize. This soon focused his attention also on hen egg-white lysozyme (Fleming, 1922), but in the tetragonal rather than the triclinic crystal form. He quickly learned to grow crystals by the method described by Alderton *et al.* (1945) and then found that precession photographs of crystals soaked in uranyl nitrate showed intensities that differed significantly from those given by the native crystals. Encouraged by these results, he asked Max Perutz whether he could join the Cambridge Laboratory, but Max, having no room in Cambridge, suggested that he write to Sir Lawrence Bragg about going to the Royal Institution. Bragg replied with an offer of a place to work on β-lactoglobulin with David Green, who had by then returned to London. RJP accepted the offer and left for London late in 1960 – after first discussing what was going on at the Royal Institution with ACTN, who had just arrived at MIT for a year's leave with Alex Rich.

Early in 1961, RJP showed Bragg his precession photographs of potential lysozyme derivatives, and Bragg enthusiastically encouraged him to continue the work, at the same time urging DCP to arrange as much support as possible. This was a characteristic response by Bragg, who was well aware that at least two other groups were already working on lysozyme, Dickerson and Steinrauf at Urbana and Pauling and Corey at Cal Tech (Corey *et al.*, 1952): competition with Pauling was a common feature of his career. In describing his reaction to Bragg's encouragement, RJP recalled Metchnikoff's view of Pasteur. 'He transferred his enthusiasm and energy to his colleagues. He never discouraged anyone by the air of scepticism so common among scientists who had attained the height of their success ... He combined with genius a vibrant soul, a profound goodness of heart.'

25.1.2. Structure analysis at 6 Å resolution

25.1.2.1. *Technical facilities*

In 1961, the Davy Faraday Laboratory was well equipped with X-ray generators. They included both conventional X-ray tubes, operating at 40 kV and 20 mA to produce copper $K\alpha$ radiation, and high-powered rotating-anode tubes that had been built in the laboratory to the design of D. A. G. Broad (patent 1956) under the direction of U. W. Arndt. We had a number of Buerger precession cameras and a Joyce–Loebl scanning densitometer, which had been used in the analysis of myoglobin (Kendrew *et al.*, 1960). In addition, we had a laboratory prototype linear diffractometer (Arndt & Phillips, 1961), which had been made in the laboratory workshop by T. H. Faulkner, and the manually operated three-circle diffractometer that had been used to make some of the measurements in the 6 Å studies of myoglobin (Kendrew *et al.*, 1958) and haemoglobin (Cullis *et al.*, 1961). The diffractometers were used with sealed X-ray tubes, since the rotating anodes were not considered to be reliable or stable enough for this purpose.

At this stage, most of the computations were done by hand, but we did have access to the University of London Ferranti MERCURY computer, usually in the middle of the night. This machine was programmed in MERCURY Autocode. The development of the early computers, their control systems and compilers mentioned in this article have been described by Lavington (1980).

Copyright © 2012 International Union of Crystallography

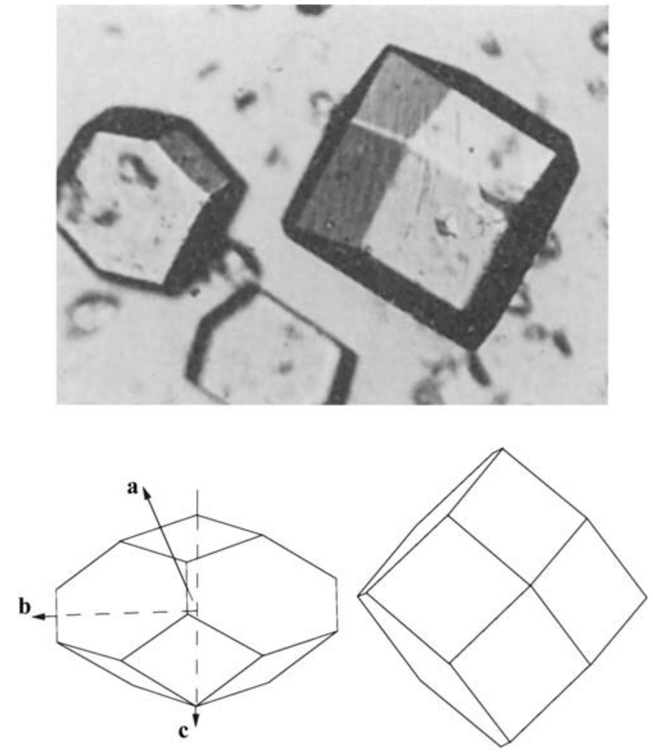


Figure 25.1.2.1
Tetragonal lysozyme crystals with well developed {110} faces (left-hand crystal) and small {110} faces (right-hand crystal).

25.1.2.2. *Lysozyme crystallization*

Tetragonal lysozyme crystals were first reported by Abraham & Robinson (1937) and the standard method of preparation was developed by Alderton *et al.* (1945); RJP used this method. Lyophilized lysozyme was obtained commercially and dissolved in distilled water at concentrations ranging from 50 to 100 mg ml^{-1}. To a volume of the lysozyme solution, an equal volume of 10% (*w*/*v*) NaCl in 0.1 *M* sodium acetate (pH 4.7) was added. About 1 to 2 ml aliquots of this mixture were pipetted into glass vials and tightly capped. Large crystals, frequently with volumes in the range 0.5 to 1 mm^3, grew overnight or during the course of a few days.

The crystals were modified bipyramids with well developed {011} faces and bounded by hexagonal {110} faces that were developed to differing extents in individual crystals (Fig. 25.1.2.1). Many crystals grew in contact with the glass and were less regular in shape. Precession photographs confirmed that the unit-cell dimensions were $a = b = 79.1$ Å and $c = 37.9$ Å, and that the space group was $P4_12_12$ or $P4_32_12$ (Corey *et al.*, 1952). Each unit cell contained eight lysozyme molecules (one per asymmetric unit), molecular weight about 14 600, together with sodium chloride solution that made up about 33.5% of the weight of the crystal (Steinrauf, 1959).

With respect to structure analysis by the method of isomorphous replacement, these two enantiomorphic space groups have the great advantage of exhibiting three independent centrosymmetric projections, on the (001), (010) and (110) planes, corresponding to the *hk*0, *h*0*l* and *hhl* reflections, respectively. As a result, 173 of the 393 reflections from planes with spacings ≥ 6 Å have heavy-atom contributions exactly in or exactly out of phase with the protein contributions to the structure factors. This property greatly facilitated the determination and refinement of heavy-atom positions in the isomorphous derivatives used in the work on lysozyme.

25.1.2.3. *Preparation of heavy-atom derivatives*

For this work by RJP, lysozyme crystals were grown in glass vials in which a 1 ml solution contained between 20 and 50 mg of lysozyme. Attempts were then made to diffuse heavy-atom compounds into the pre-formed crystals. Stoichiometric amounts of heavy-atom compounds, such as K_2PtCl_4, $UO_2(NO_3)_2$, *p*-chloromercuribenzene sulfonate (PCMBS), *p*-chloromercuribenzoate (PCMB) and K_2HgI_4, were added to these vials, and precession photographs were taken of the crystals from each vial. The precession photographs of the putative derivative crystals were compared visually with those from the native protein by superimposing them on a light box. This showed immediately whether the cell dimensions had changed and if there were any significant changes in intensity. If these photographs showed no changes in intensity, then the amount of heavy-atom reagent was increased, and this process was continued until the crystals either showed intensity changes or disintegrated. When intensity changes were detected, the effect of increasing the concentration of heavy atom was explored with the object of establishing the optimum conditions for the preparation of derivatives with high occupancy of a small number of sites. In this way it was sometimes possible to follow decreases and increases of the intensities of weak reflections that went through zero with increasing concentration, and this indicated a reversal of the signs of the reflections from the native and derivative crystals.

In accordance with the example provided by the work on myoglobin, 9° precession photographs, which provide data to a resolution just beyond 6 Å, were used for these trials, and attention was concentrated on the [001] and [100] zones of reflections.

When these exploratory studies had produced a promising derivative, further precession photographs were taken for use in intensity measurements. For this purpose, two Ilford X-ray films were placed one behind the other in the camera cassette and they were exposed for about 24 h to Cu $K\alpha$ radiation from sealed X-ray tubes or for about 4 h to the same radiation from a rotating-anode tube. A second exposure with two films in the cassette was made for about 4 h with the sealed tube or 1 h with the rotating-anode tube in order to cover the full range of intensities that had to be measured. The intensity measurements were performed on the Joyce–Loebl recording densitometer, which scanned each row of reflections automatically but had to be moved manually from row to row. The heights above background of the peaks on the densitometer traces were measured with a ruler and these measurements provided the basic intensity data. Intensities recorded on the two films of each pair were brought to the same scale by calculation and application of a film transmission factor (usually between 2.5 and 3.0) and the corresponding factor relating films exposed for different lengths of time was obtained similarly. Weighted mean intensities were then calculated for each of the reflections. Lorentz–polarization (Lp) factor corrections were derived from a plot of the Lp factor against ($\sin^2\theta/\lambda^2$), without consideration of asymmetric effects (Waser, 1951), and these factors were manually applied to the observed intensities to provide structure-factor measurements on an arbitrary scale.

The structure factors of the heavy-atom derivative crystals, $|F_{HP}|$, were scaled to those of the native protein, $|F_P|$, by a factor K, derived from the equation

$$K^2 \sum |F_{HP}|^2 = \sum |F_P|^2 S,$$

where S was a parameter, between 1.0 and 1.10, that depended on the assumed heavy-atom occupancy and was refined later in the process. Values of $\Delta F = K|F_{HP}| - |F_P|$ were then calculated for use in the determination of the heavy-atom positions.

25.1.2.4. Determination of heavy-atom positions

The high symmetry of the space group was greatly to our advantage, since the heavy-atom positions could be determined from difference-Patterson projection on the (100), (110) and (001) planes. In principle, all three coordinates of a heavy atom can be determined from projections on (100) or (110) alone. In practice, however, it was more straightforward to begin with the interpretation of the simpler projection on (001) before determining the z coordinate of the heavy atom from one of the other projections.

Apart from the effect of cross-over terms, which were sometimes detected as indicated above, these maps are effectively true Patterson maps of the heavy-atom structures of the derivatives. These difference-Patterson maps were calculated on the MERCURY computer and even from time to time by the use of Beevers–Lipson strips – a less demanding task than it might appear, since only about 80 *hk*0 and 60 0*kl* reflections are included within the 6 Å limit.

25.1.2.4.1. The mercuri-iodide (K_2HgI_4) derivative

After trying several levels of substitution, RJP used the K_2HgI_4 salt at a molar concentration eight to ten times that of lysozyme. Based on the fact that lysozyme contains two methionine residues per molecule, and in keeping with a suggestion of Bluhm *et al.* (1958), RJP was expecting to see two heavy-atom sites,[1] but the *hk*0 difference-Patterson map was interpretable in terms of a single site of substitution. This site, however, was very close to the crystallographic twofold axis that runs along a diagonal in the [001] projection of the unit cell, and it proved necessary to correct the details of the first interpretation when phase information became available from other derivatives. It then appeared that there was one HgI_4^{2-} (or HgI_3^-) on the twofold axis between two protein molecules, but that it was best modelled by two closely spaced sites to allow for the elongated shape of the group (see Table 25.1.2.1). Several other heavy-atom salts, including K_2HgBr_4, K_2PtBr_4 and K_2AuCl_4, gave derivatives in which the heavy atom was attached to the same site as K_2HgI_4, and consequently seemed not to provide useful additional phase information.

25.1.2.4.2. The palladium chloride (K_2PdCl_4) derivative

An attempt to use K_2PtCl_4 to produce a useful derivative gave disordered crystals, but a substitute for it was found by soaking crystals in K_2PdCl_4 at a molar ratio of 3:1 relative to lysozyme. Despite the relatively light Pd atom, this derivative gave an easily interpretable difference-Patterson map (see Fig. 25.1.2.2) that yielded very good R factors,

$$R = \textstyle\sum ||F_{HP}| - |F_P|| - |F_H(\text{calc})| / \sum |F_H(\text{calc})|,$$

where the summations are over centric reflections only.

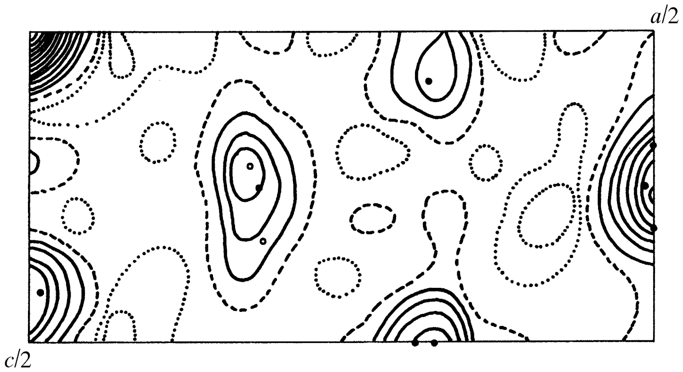


Figure 25.1.2.2
Difference-Patterson *h*0*l* projection map for the derivative obtained with K_2PdCl_4. The ends of heavy-atom double-weight (filled circles) and single-weight (open circles) vectors are shown (R. J. Poljak, unpublished material).

25.1.2.4.3. The o-mercurihydroxytoluene p-sulfonate (MHTS) derivative

The *hk*0 difference-Patterson map of the *p*-mercuribenzene sulfonate (PCMBS) derivative was interpretable in terms of a single site of substitution at 8 Å resolution, but it was not useful beyond about 8 Å because of lack of isomorphism. RJP and RHF then explored the usefulness of MHTS as a derivative. This compound had been specially synthesized by JWHO in the hope that a small rearrangement of groups present in PCMBS would lead to an isomorphous derivative. Happily, this strategy worked, and MHTS gave a useful isomorphous derivative in which the major site overlapped that of PCMBS (Fig. 25.1.2.3).

25.1.2.4.4. Other potential derivatives

As is usual in protein work, RJP tried many other heavy-atom compounds (Poljak, 1963), but none gave useful results. In particular, a uranyl derivative, obtained by the use of UO_2NO_3,

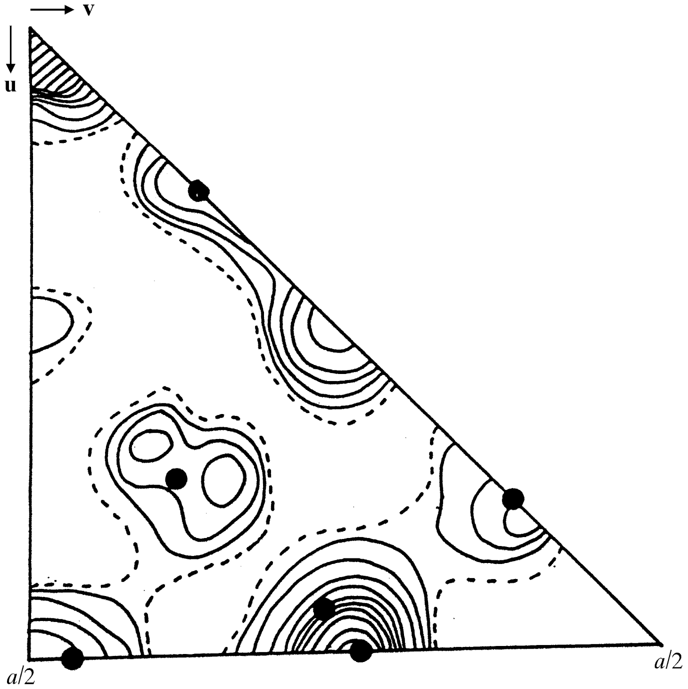


Figure 25.1.2.3
Difference-Patterson *hk*0 projection for the derivative obtained with MHTS. The large peak at $\frac{1}{4}$, $\frac{1}{4}$ is not explained by this solution (Fenn, 1964).

---

[1] In order to study the nature of the ligand formed from K_2HgI_4, RHF and DCP studied the structures of two compounds, $(CH_3)_3S{\cdot}HgI_3$ and $[(CH_3)_3S]_2{\cdot}HgI_4$ (Fenn *et al.*, 1963; Fenn, 1964), which were prepared for this purpose in the laboratory by JWHO. These studies showed that HgI_3 and HgI_4 are, respectively, planar trigonal and tetrahedral in configuration. Meanwhile, Dr Helen Scouloudi was examining the nature of the K_2HgI_4 derivative of seal myoglobin (Scouloudi, 1965) and showed that the ligand was HgI_3^- and not associated with methionine.

gave difference-Patterson maps that were difficult to interpret, and it was not taken further at this stage.

25.1.2.5. Refinement of heavy-atom parameters

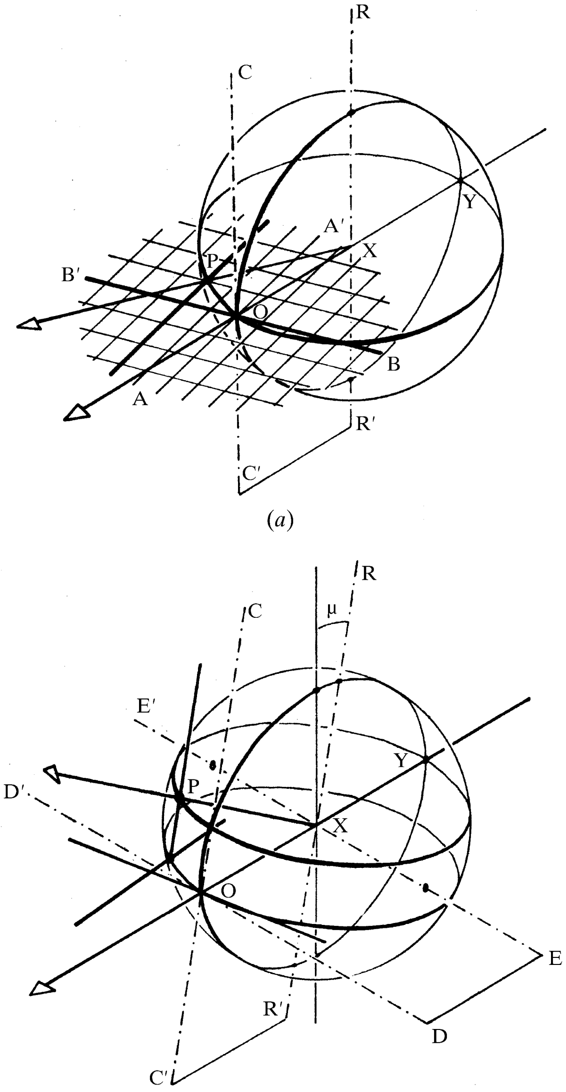
Refinement of the heavy-atom parameters was first performed by the use of Rollett's (1961) least-squares program on the MERCURY computer, using the $|\Delta F|$ values as structure amplitudes. This procedure gave satisfactory results for the K_2HgI_4, K_2PdCl_4 and MHTS derivatives described above, and they were used, therefore, in an attempt to determine the structure of the protein to 6 Å resolution in three dimensions.

25.1.2.6. Analysis in three dimensions

25.1.2.6.1. X-ray intensity measurements

We had three options for the collection of three-dimensional data. First, we could have used precession photographs and densitometry, as in the study of myoglobin (Kendrew *et al.*, 1960). Second, the manually controlled three-circle diffractometer used to make some of the measurements in the 6 Å study of myoglobin (Kendrew *et al.*, 1958) and haemoglobin (Cullis *et al.*, 1962) was available and, third, there was the prototype linear diffractometer (Arndt & Phillips, 1961). We chose the last option, since CCFB, RHF and DCP were well experienced in using the instrument, and it offered the opportunity of measuring the protein reflections automatically and in a relatively short time compared with the other methods. Before he went on leave to MIT, ACTN had written a computer program for the University of London MERCURY computer to process the diffractometer data (North, 1964) and, on his return in September 1961, he readily accepted an invitation to join the team to continue with this and other related aspects of the work.

The design of the linear diffractometer was based directly on the reciprocal-lattice representation of the genesis of X-ray reflections. The principle is illustrated in Figs. 25.1.2.4(*a*) and (*b*), which show the familiar Ewald construction. YXO represents the direction of the incident X-ray beam with X the centre of the Ewald sphere and O the origin of the reciprocal lattice. A′OA, B′OB and C′OC are the principal axes of the reciprocal lattice, here assumed to be orthogonal. XP is the direction of the reflected X-ray beam corresponding to the reciprocal-lattice point P, which lies on the surface of the sphere of reflection. The reciprocal lattice can be rotated about the axis C′OC, and this axis can be inclined to the direction of the incident X-ray beam by rotation about the axis D′OD, which is perpendicular to the incident beam.



(*a*)

(*b*)

Figure 25.1.2.4

Reciprocal-space diagrams showing the direction of the incident X-ray beam, the Ewald sphere and the genesis of a reflection (*a*) in an equatorial plane and (*b*) in the equi-inclination setting. Principal reciprocal-lattice directions are shown as thick lines. They also represent the slides in the diffractometer. The rotation of the diffractometer slide system about the axis C′OC is coupled to the rotation of the crystal about the axis R′XR by gears, pulleys and steel tapes. The counter arm of the diffractometer is represented by the fixed link XP = XO. Reproduced with permission from Arndt & Phillips (1961).

The linear diffractometer was simply a mechanical version of this diagram. The reciprocal lattice was represented by three slides, A, B and C, which were parallel, respectively, to A′OA, B′OB and C′OC. They were mounted to rotate about the axis C′O and arranged so that the saddle P could be set to any position in space within the coordinate system that they defined. This saddle P was connected to the point X by means of a link of fixed length, XP = XO, corresponding to the radius of the sphere of reflection. The link XP always lay along the direction of the reflected X-ray beam and thus became the counter arm of the diffractometer. The crystal was mounted at X for rotation about the axis R′XR (independent of the link XP, which pivoted about an independent coaxial bearing at X). The rotation of the crystal about this axis was coupled by means of gears, pulleys and steel tapes to the rotation of the slide system about the axis C′OC. The axes R′XR and C′OC, held parallel by means of parallel linkages, could be tilted with respect to the incident X-ray beam by rotation about the axes D′OD, E′XE, as shown in Fig. 25.1.2.4(*b*).

The scale of the instrument clearly depended only on the length chosen for XP = XO. In the instruments used in the lysozyme work, this length, which is equivalent to one reciprocal-lattice unit, was five inches. The position of the saddle P on the three slides was controlled by means of lead screws, all of which were cut with 20 turns per inch. Hence the counters, which indicated revolutions and fractions of a revolution of the lead screws, read directly in decimal divisions of reciprocal-lattice units. The screws in slides A and B were driven by means of synchro-receiver motors, forming a synchro link with corresponding transmitters in the control panel. Slide C was set manually, together with the inclination angle μ, for the measurement of upper-level reflections in the Weissenberg equi-inclination mode, Fig. 25.1.2.4(*b*).

The coupling of the rotations of the crystal and reciprocal lattice about the axes R′XR and C′OC, respectively, was interrupted by two ancillary mechanisms. The first simply allowed for independent rotation of the crystal with respect to the slide system and was used for setting the reciprocal-lattice axes in the equatorial plane parallel to slides A and B, and for any fine adjustment of the crystal rotation that might be necessary during the measurement procedure. The second interruption consisted of a mechanism for oscillating the crystal about the position for any reflection while X-ray intensity measurements were made. This oscillation mechanism (Arndt & Phillips, 1961) rotated with the crystal as the diffractometer was being set to a reflection position, and then controlled the independent motion of the crystal for the measurement of the integrated intensity of the reflection. The crystal remained stationary at a given angular setting for time *t*, was rotated at a uniform rate over a predetermined angular range for a time 2*t*, remained stationary at the final angular setting for a further time *t*, and then returned quickly to its original setting. The correct setting for the reflection peak was at the midpoint of the rotation, which might be set to be through any angle from 1 to 5°. For initial adjustments, the motor could be arrested at this midpoint by means of a micro-switch operated by a switching disc rotating with the cam. This disc otherwise actuated contacts that started and stopped the intensity measurements.

The X-ray intensities were measured with a side-window xenon-filled proportional counter made by 20th Century Electronics, together with associated amplifiers and pulse-counting circuits (Arndt & Riley, 1952). The proportional counter had a high quantum efficiency for the measurement of Cu $K\alpha$ radiation (about 80%) and, when the operating potential and pulse-height discriminating circuits were carefully set, it provided useful discrimination against radiation of other wavelengths. The output from the proportional counter and its associated circuitry was fed directly to a teleprinter, which gave both a plain-language print out and a five-hole punched paper tape for input to the computer. Each count was provided with a check digit derived by a 'ring-of-three' circuit, wired in parallel with the main electronic counter. During data processing, the check digit was compared with the count *modulo* 3: inequality of the two numbers was taken to indicate an error in the counting circuit.

Three counts were made: the first was a background count n_1, made while the crystal was stationary on one side of the reflection position; the second was an integrated intensity count N, accumulated as the crystal rotated through the reflection; and the third was a further background count n_2. The background-corrected integrated intensity of the reflection was taken to be

$$N_o = N - (n_1 + n_2).$$

At this stage of the work, measurements were made of one reciprocal-lattice level at a time in the equi-inclination mode that has no blind region at the centre. In each level, the diffractometer was driven to a reflection *hkl* (for example) at the limit of resolution of the data set to be collected. The diffractometer then moved in a series of equal steps along the scanning slide. At the end of each step, the oscillation mechanism took control for the measurement of intensity and background. After each measurement, a further step was taken on the scanning slide, and the process continued until a limit switch, set to the required resolution limit, was reached. The diffractometer then completed the current translation, measured the last reflection in that row, and then moved one step on the stepping slide to the next parallel row. This row was scanned, in the opposite direction, until the limit switch was reached again. In this way, the whole of a reciprocal-lattice level could be measured. In order to change to another level, the vertical slide C and the inclination angle μ had to be manually adjusted.

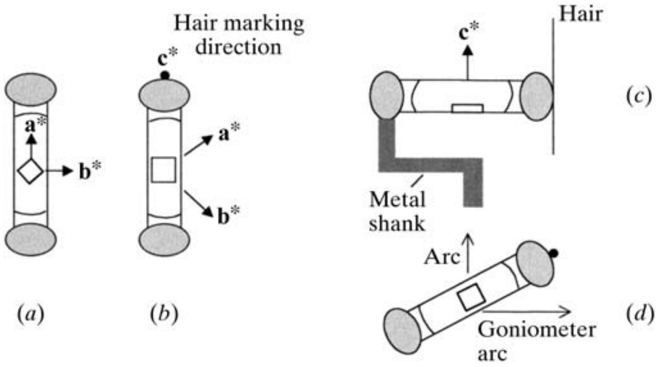


Figure 25.1.2.5
Crystal mounting. (*a*) Rotation about the a^* axis; (*b*) rotation about [110], preliminary to c^* mounting; (*c*) rotation about the c^* axis (elevation); and (*d*) rotation about the c^* axis (plan).

This account ignores two difficulties, one inherent in the design of the diffractometer, and the other specific to the lysozyme crystals. First, the instrument required a good deal of supervision, since it did not set itself very well for the measurement of low-angle reflections. Second, the crystals were not easily mounted so that the *c* axis, the most convenient axis for efficient data collection, since it is perpendicular to the most densely populated reciprocal-lattice planes, coincided with the crystal-rotation axis of the diffractometer.

The first problem was overcome by efficient teamwork and was much eased by the fact that RHF assumed responsibility for the MHTS derivative as part of her PhD work; the second was solved by making most of the measurements from crystals mounted to rotate about the [100] axis. These crystals were oriented so that the b^* and c^* axes were parallel to the horizontal slides of the diffractometer, and the measurements were made in levels of constant *H* by scanning along rows parallel to b^* and stepping to adjacent rows along c^*. A number of reflections could not be measured in this way, however, because reflections near the a^* axis were too broad to measure, particularly in the upper reciprocal-lattice levels. This difficulty was overcome by mounting some crystals with the *c* axis of the tetragonal crystals perpendicular to the length of the capillary tube, with the [110] axis parallel to the tube. These specimens were then mounted on a right-angled yoke so that the capillary tube was perpendicular rather than parallel to the goniometer axis (Fig. 25.1.2.5).

Given the morphology of the crystals, with an essentially square habit bounded by {110} faces (Figs. 25.1.2.1 and 25.1.2.5), all the reflections in the *hkl* octant could be measured in levels with constant *L* values without inclining the capillaries by more than about 40° to the X-ray beam. The horizontal slides of the diffractometer were set to be parallel to the a^* and b^* axes of the crystal. Some quadrants of *hkL* reciprocal-lattice levels were then scanned along rows parallel to the a^* axis and stepped along the b^* axis. Enough measurements were made in this mode to cover the 'blind' region in the *Hkl* levels and provide an appropriate number of intersecting levels for scaling all the measurements into a consistent set.

Care was taken during all these measurements to index the reflections in a right-handed system of axes. Given the transparent relationship between the slide system of the diffractometer and the crystal geometry, this was easily accomplished, and it was necessary for the subsequent use of anomalous scattering (Bijvoet, 1954) in the phase determination.

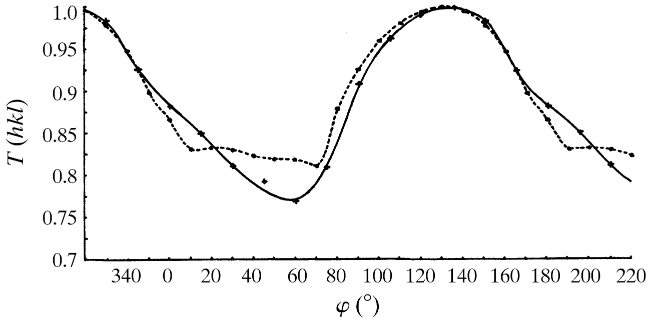


Figure 25.1.2.6
Absorption curve. Variation of relative transmission $T(hkl) = I(\varphi_{hkl})/I_{max}(\varphi)$ [$= 1/A(hkl)$] with rotation angle φ for the 200 reflection and the crystal rotating about the normal to (200). Solid line: measured curve; broken line: calculated curve, neglecting effect of mother liquor and capillary. Reproduced with permission from North *et al.* (1968).

During the measurements from the native and derivative crystals mounted for rotation about [100], the variation in peak intensity of the 200 reflection with φ, the angle of rotation about the axis C′OC (Fig. 25.1.2.4*a*), was also recorded (Fig. 25.1.2.6). (200 is the lowest-order reflection available for this purpose in this space group.) These records were then used in the data-processing stage to correct the measurements for absorption by the method described by Furnas (1957). According to this method, the absorption suffered by the incident and reflected beams for any reflection *hkl* is indicated by the relative intensity of 200 when the X-ray beams are parallel to the reflecting planes *hkl*. This is a fair approximation for reflections at low angles. This method could not be used, however, for the measurements made from crystals mounted to rotate about the [001] axes, since a full rotation about this axis was not possible. These measurements were not corrected for absorption errors.

Finally, the diffractometer was reset manually at regular intervals during data collection to measure the intensities of a number of reference reflections. These measurements were used to monitor the stability of the system and the extent of irradiation damage to the crystal, and they were also recorded on the paper-tape output for analysis by the data-processing program. An attempt was made to minimize irradiation damage by using a shutter to expose the crystal only during the measuring cycle.

25.1.2.6.2. *Data processing*

The measurements N, n_1 and n_2, together with the indices of the reflections, *hkl*, were all printed out in plain language on a teleprinter and punched in paper tape for direct transfer to a computer (Fig. 25.1.2.7). The plain-language record was important during measurement of the low-angle reflections, when the diffractometer had to be adjusted by hand. Not all imperfections in the measurements were easily spotted at this stage, however, and ACTN's data-processing program (North, 1964) therefore incorporated systematic checks on the quality of the measurements.

The program checked for the following contingencies:

(1) malfunction of the diffractometer-output mechanism leading to the paper tape being an inaccurate record of the measurements, generally because the tape punch had failed to perforate the tape or had 'stuttered';
(2) errors by the pulse counters, detected by the 'ring-of-three' circuit;
(3) peak counting rate so high that counting-loss errors were appreciable;
(4) count on reflection not significantly above background;
(5) failure of diffractometer to set crystal or counter correctly; and
(6) gradual drift in the experimental parameters, including movement of the crystal within its mounting and irradiation damage to the crystal.

```
*02+13+00+
:000234 002500 000284

*01+13+00+
:000260 001660 000279

*00-13+00+
:000181 000522 000212

*00-12+00+
:000225 000651 000236

*01+12+00+
:000327 005860 000362

*02+12+00+
:000281 006717 000260

*03+12+00+
:000236 000507 000240

*04+12+00+
:000225 000488 000241

*05+12+00+
:000244 003763 000263
```

(*a*)

```
*    7     2
 0   4912   98
 1   4836   97
 2    276   33
 3  12710  127
 4   1216   49
 5    802   40
 6    398   32
 7   1551   47
 9    935   37
10   1257   38

     8     2
 0   9855   99
 1   1450   44
 2   6117   61
 3   4733   95
 4     61   26
 5    395   32
 6   1304   39
 7    129   28
 8    183   27
 9   1725   52
```

(*b*)

Figure 25.1.2.7
Typical output from the linear diffractometer. (*a*) Indices *h*, *k*, *l* followed by background (n_1), peak (N), background (n_2) counts. (*b*) Listing ready for the next stage in data processing with indices * *h k* followed by *l*, background corrected peak and standard deviation. Reproduced with permission from North (1964). Copyright (1964) Institute of Physics.

These checks were made while the diffractometer tape was being read into the computer, and a monitor output was produced simultaneously, as shown in Fig. 25.1.2.8. The checks

```
PBDN 3/2
TLY  PDBR 4   D 2 * 1/5     K = 12:14:16      24/2/64
 01    01    14  :=  SD  966
 01    00    14  PE 2
 01   -01    14  :-  SD  999  :=  SD  156
 01   -04    14  :-  BG  :+  BG
 01   -13    14  :=  N-  SD  578  BG
 01   -13    14  :-  N-  :=  N-  BG
 01   -13    14  :-  SD  391  :=  N-  SD  234
 01   -14    14  :-  SD  132  :+  SD  216
 01   -15    14  :=  SD  813
 00   -15    14  :-  BG  :=  BG  :+  BG
 00   -12    14  :=  BG
 00   -10    14  :+  BG
 00   -10    14  :=  BG
 00   -04    14  :=  BG
 00   -04    14  :+  BG
 00    04    14  :-  BG  :=  BG  :+  BG/

TOTAL  PE1    PE2    PE*    SD    N-    BG    N>H
 42     0      1      0      9     4    15      0
```

Figure 25.1.2.8
Format of monitor output in which the computer lists reflections that fail the tests for format or significance. PE1 signifies punching error, indices; PE2, punching error, measurements; PE3, failure of electronic check on counting circuits; SD, standard deviation greater than set limit; N-, net count negative; BG, backgrounds significantly different; N > H, gross counts exceed counting-loss limit. This output was from the version of the program designed to be used with the diffractometer fitted with three counters. The symbols: −, =, + refer, respectively, to reflections measured by the lower, central and upper counters. Reproduced with permission from Arndt *et al.* (1964). Copyright (1964) Institute of Physics.

depended in large part on the fact that the significance of an intensity measurement may be assessed in terms of counting statistics. The standard deviation of a background-corrected count, N_o $(= N - n_1 - n_2)$, is given by $\sigma^2(N_o) = N + n_1 + n_2$, and the ratio $\sigma(N_o)/N_o$ may be taken as an indication of the significance of the measurement. Measurements were rejected when this ratio exceeded unity. N_o might then have been taken as zero but, following Hamilton (1955), we considered it preferable to replace N_o by a fraction (0.33 for centric and 0.5 for acentric reflections) of the minimum background-corrected count that we should have considered acceptable. Reflections were treated in the same way whether the net count N_o was positive or negative, but measurements were rejected if N_o was negative and $|N_o| < \sigma(N_o)$.

Mis-setting of the crystal was frequently revealed by marked inequality of the background counts. Measurements were therefore rejected if the difference between the two backgrounds exceeded three or four standard deviations, that is if $(n_1 - n_2)^2 > b^2(n_1 + n_2)$, where b is the appropriate constant.

After monitoring the quality of the data in this way, the program proceeded: (i) to extract background-corrected counts; (ii) to apply a correction for irradiation damage derived from any systematic variation in the intensities of the reference reflections; (iii) to sort the reflections into a specified sequence of indices; (iv) to apply Lorentz–polarization factors; and (v) to apply absorption corrections (the data for which were read separately from a specially prepared punched tape, Fig. 25.1.2.6). The outputs from this program comprised data sets from a number of individual crystals of the native protein and the three derivatives. The scale factors needed to bring the measurements from the individual crystals of each crystal species to a common scale were determined from the intensities in common rows by the use of a program written by Rollett (Rollett & Sparks, 1960). The scaling factors were applied and the data merged, a weighted-average intensity being determined when more than one estimation was available.

25.1.2.6.3. *The absolute scale of the intensities*

An attempt was made to determine the absolute scale of the measured intensities by comparison with the intensities diffracted by anthracene, a small organic crystal of known structure. This method had worked well in a determination of the absolute scale for seal myoglobin (Scouloudi, 1960), but it did not give a satisfactory result with lysozyme, mainly because of the difficulty of measuring the crystal volumes precisely enough. Accordingly, we used Wilson's (1942) method to provide an estimate of the absolute scale of the intensities, knowing very well that it does not give an accurate estimate for protein data, especially at low resolution. Nevertheless, this scale gave reasonable values for the occupancies of the heavy-atom sites.

25.1.2.6.4. *Re-assessment of heavy-atom derivatives*

Given the three-dimensional data to 6 Å resolution for the native crystals and the three derivatives, it was next possible to calculate three-dimensional difference Patterson maps for the derivatives using the terms

$$|\Delta F|^2 = ||F_{PH}| - |F_P||^2$$

as coefficients in the Fourier series. This synthesis, which is now well known in protein-structure analysis, gives a modified Patterson of the heavy-atom structure in which the heavy-atom vectors appear at reduced weight in a complex background (Blow, 1958; Phillips, 1966). Nevertheless, the $|\Delta F|^2$ maps for $PdCl_4$ and MHTS were readily interpreted in terms of single heavy-atom substitutions, particularly because the vectors involved were all confined to defined Harker sections (Fig. 25.1.2.9).

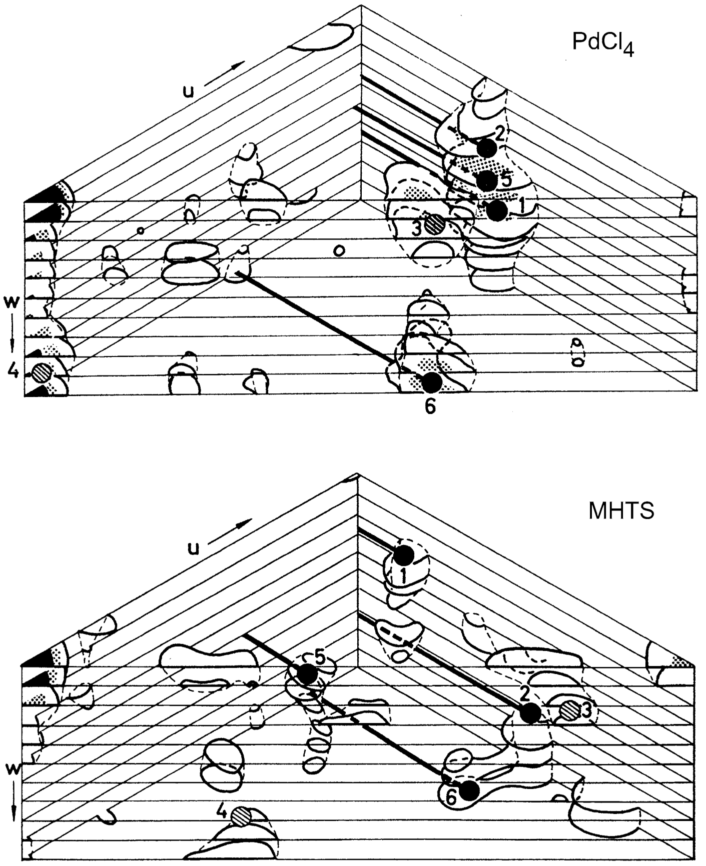


Figure 25.1.2.9
Three-dimensional $|\Delta F|^2$ syntheses for the $PdCl_4$ and MHTS derivatives. The positions of peaks in the Harker sections are marked and numbered (Fenn, 1964).

The map of the HgI_3^- derivative was not so satisfactory, and this led to the discovery, during refinement of the heavy-atom parameters, that the mercury occupancy declined as a function of irradiation time. Consequently, the HgI_3^- data from two individual crystals were treated separately during the remaining stages of the analysis.

The least-squares refinement mentioned in Section 25.1.2.4 was not wholly satisfactory in that it included no provision for refining heavy-atom occupancy. Accordingly, DCP – with some help from ACTN – wrote a computer program for the MERCURY computer based on Hart's (1961) method, in which the heavy-atom positions, occupancies (O) and temperature factors (B) were refined simultaneously together with the scale factors (S) between heavy-atom and native structure factors. All the centric reflections from the [100], [110] and [001] zones were included in this refinement.

The quantity minimized for each derivative was

$$R' = \textstyle\sum(||F_{PH}| - |F_P|| - |F_H|)^2,$$

where F_H is the calculated heavy-atom contribution to the derivative structure factor. The method involves calculating R' for each parameter at its current value p_n and at values $p_n \pm \Delta p_n$ and $p_n \pm 2\Delta p_n$, where each of the parameters is shifted in turn, the shifts having been specified, while the other parameters are at the unshifted value. Thus for each of the parameters (x, y, z, O, B and

Table 25.1.2.1

Heavy-atom parameters used in the final phase calculation for the lysozyme structure

E is the average difference between observed and calculated heavy-atom changes of centric reflections (electrons); R is the reliability index for observed and calculated heavy-atom changes of centric reflections and R' is the Kraut (Kraut *et al.*, 1962) agreement index for all reflections.

			Mercuri-iodide			
	$PdCl_4$	MHTS	Crystal 1†		Crystal 2‡	
x	0.147	0.218	0.131	0.178	0.125	0.178
y	0.841	0.620	0.869	0.822	0.875	0.822
z	0.963	0.054	0.250	0.250	0.250	0.250
Occupancy (e)	72	47	74	74	72	72
B (Å^2)	52	−18	38	38	111	111
E (e)	57	66	66		55	
R (%)	35	49	42		39	
R' (%)	10.7	11.6	11.6		12.2	

† Centric 0, 1, 2*kl* reflections only. ‡ *hk*0 reflections only.

S), four values of R' are obtained for the shifted values plus the value of R' for the unshifted parameters, the latter being denoted ε_u.

Let the minimum value of R' from all the calculations be $\varepsilon_{\min}$ and for the parameter p_n, the minimum of the list of five values of R', be ε_{qn}, corresponding to the value q_n of p_n. Then, according to the method of steepest descents, the shift to be applied to the parameter p_n is

$$(q_n - p_n)(\varepsilon_u - \varepsilon_{qn})/(\varepsilon_u - \varepsilon_{\min}).$$

If $\varepsilon_{qn} = \varepsilon_u$, that is, if the unshifted value of the parameter gave the minimum value of R', then the shift was divided by 4 for the next cycle. Otherwise the shift was kept constant. Thus the new parameters and shifts were determined for the next cycle of refinement, and the process was repeated until convergence.

This program worked well, and RJP, who was reading *Candide* at the time, named it *Pangloss* – it gave the best possible values for the heavy-atom parameters. These values, which include two separate sets for the HgI_3^- derivative, are shown in Table 25.1.2.1. At this stage, an important check was carried out. The coordinates of the heavy-atom site in each derivative were referred to an origin at the junction of a twofold axis and a twofold screw axis. However, there are four such intersections in the unit cell and, in order to ensure that the same origin had been chosen for each derivative, the sign predictions for the centric reflections from each derivative – which were checked by hand throughout this exploratory stage – were compared. They agreed well, thus establishing that the choice of origin was the same for each derivative.

25.1.2.7. Phase determination at 6 Å resolution

These parameters were used to determine the phases of the protein reflections. A proportion of these phases were first determined by the graphical method suggested by Harker (1956), which had been used in the 6 Å stage of sperm-whale myoglobin (Kendrew *et al.*, 1958). We treated the process as a group activity in which different individuals took responsibility for reading out the various Harker components [F_P, F_{PH}, F_H(calc) *etc.*] for each reflection in turn. This was a useful bonding exercise and improved our familiarity with the rather cosmopolitan accents in use within the group. Scientifically, it was also an encouraging experience since it showed that reasonably consistent results could be obtained from the three different derivatives, *and* that the anomalous-scattering measurements were capable of making a significant contribution to the phase determination. It soon became clear that the most significant anomalous contributions had to be included in such a way as to retard the phase of the derivative structure factor if these indications were to agree with the phase predictions derived from the isomorphous differences alone (Fig. 25.1.2.10).

The implication of this observation was that the space group is $P4_32_12$ rather than $P4_12_12$, that is, the fourfold screw axis is left handed.

These results encouraged us to go ahead with the computer calculation of phases by the phase probability method applied to the isomorphous differences (Blow & Crick, 1959; Dickerson *et al.*, 1961) and to the anomalous-scattering differences (Blow & Rossmann, 1961). This program was written by ACTN and it was used first to confirm the space-group identification. Overall, the mean figure of merit obtained with the anomalous contribution consistent with $P4_32_12$ was somewhat higher than the alternative, though to a lesser extent than we had anticipated. This observation led at a later stage to reconsideration of the way in which the anomalous scattering was incorporated in the phase determination (see Section 25.1.3.9).

The quality of the phase determination was indicated by the figure of merit. For the acentric reflections, this was 0.79, an encouraging result since it compared favourably with that obtained in the low-resolution study of haemoglobin (Perutz *et al.*, 1960). The mean figure of merit for the centric reflections, on the other hand, was 0.95, so that the overall value was 0.86. A check on the sign predictions for centric reflections showed that the three derivatives gave satisfyingly similar results.

These phases were also used to calculate difference-Fourier maps, showing the heavy atoms in the derivatives by the use of coefficients

$$|\Delta F| = ||F_{PH}| - |F_P||$$

associated with the protein phases and weighted by the figures of merit of each phase determination. These difference maps in three dimensions are shown in Fig. 25.1.2.11.

They revealed the presence of small subsidiary sites in these two derivatives, but these minor sites were not taken into account.

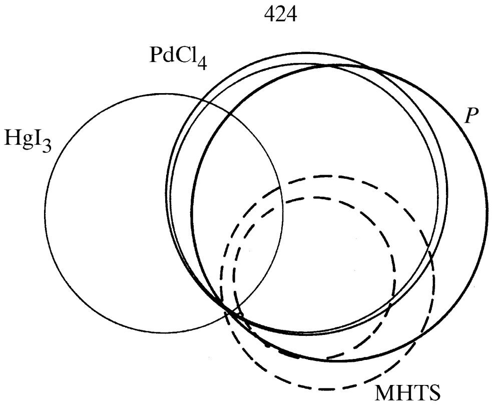


Figure 25.1.2.10

Phase determination for reflection 424. In this Harker diagram, the heavily traced protein circle (with radius $|F_P|$) is labelled P. The circles with radii $|F_{HP}|$ obtained with $PdCl_4$, MHTS and HgI_3 are shown. Values of the anomalous-scattering pairs *hkl* and *khl* were used for $PdCl_4$ and MHTS. The position of the protein vector, weighted by its figure of merit, is shown as a small open circle (R. J. Poljak, unpublished results).

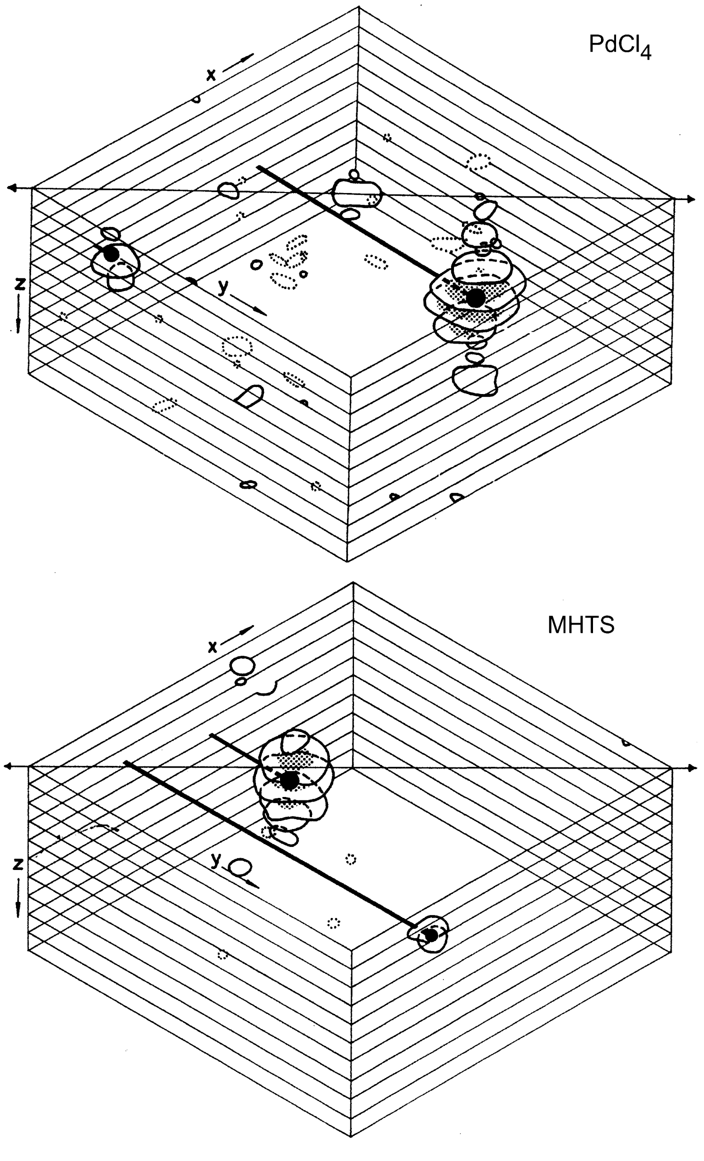


Figure 25.1.2.11
Electron-density difference syntheses showing the main sites of substitution and subsidiary sites with low occupancy in the $PdCl_4$ and MHTS derivatives (Fenn, 1964).

25.1.2.8. *The electron-density map of lysozyme at 6 Å resolution*

The electron-density distribution was calculated on the MERCURY computer, by means of a program written by Owen Mills that was in general use at the time, with structure amplitudes weighted by the figures of merit so as to give the 'best' Fourier (Blow & Crick, 1959). The map was contoured by hand and plotted on clear plastic (Perspex; Plexiglass in the USA) sheets (Fig. 25.1.2.12).

The first objective in studying this map was to determine the boundary of a single molecule. Comparison of the unit cells of various crystal forms of lysozyme (Steinrauf, 1959) suggested that in tetragonal lysozyme, the molecule occupies the full length of the *c* axis and, on average, one-eighth of the *ab* plane. On this basis, the map of Fig. 25.1.2.12 includes the whole of the *c* axis and a sufficient area of each section to ensure that one whole molecule is included in addition to parts of neighbouring molecules. Only contours indicating where the electron density is greater than average are included.

Some featureless regions of average electron density (0.4 e $Å^{-3}$) were immediately apparent. The most marked of these was around the twofold screw axis parallel to **c**. This axis is intersected at intervals of 9.5 Å by twofold rotation axes, and packing considerations, therefore, made it impossible for substantial parts of the molecule to penetrate into the neighbourhood. Similarly, the immediate vicinity of the fourfold screw axis was also without significant features and was clearly a region of intermolecular space. The twofold rotation axes also helped to determine the boundary, particularly where relatively high density approached or intersected them, as it did in two places. It was clear that such regions must represent close contacts or bridges between adjacent molecules.

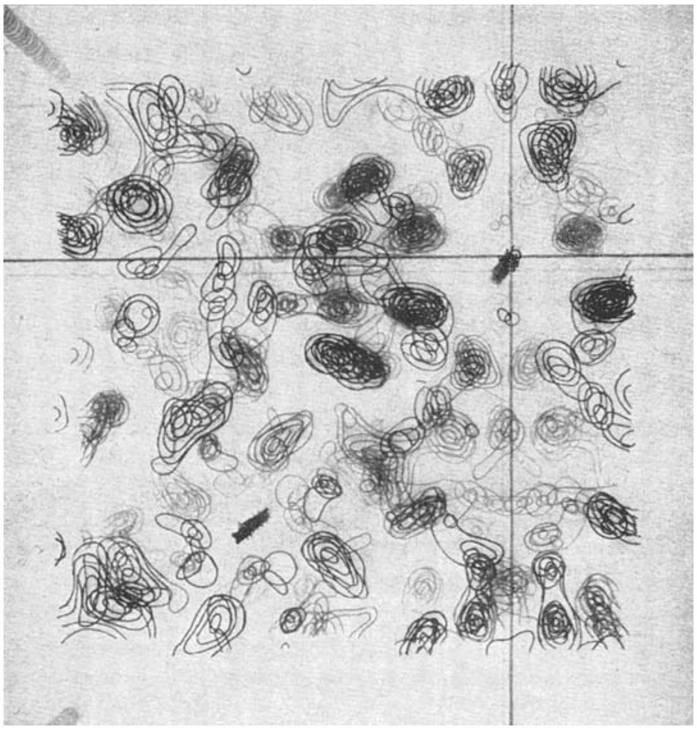


Figure 25.1.2.12
Electron-density distribution in lysozyme at 6 Å resolution viewed parallel to the *c* axis. The horizontal and vertical lines represent the twofold rotation axes and intersect the twofold screw axis, upper right of centre. The fourfold screw axis is at the lower left of centre. The contour interval is 0.07 e $Å^{-3}$, the lowest heavy contour being at 0.6 e $Å^{-3}$. The absolute scale is approximate. Reproduced with permission from *Nature* (Blake *et al.*, 1962). Copyright (1962) Macmillan Magazines Limited.

Following the example of the haemoglobin study (Perutz *et al.*, 1960), we decided at this stage to make a balsa-wood model of the electron density to help visualize the molecule. Instead of producing a stack of sections through the molecule, however, CCFB devised a way of shaping the sections to make a smooth model of the volume occupied by electron density greater than about 0.53 e $Å^{-3}$. The result is shown in Fig. 25.1.2.13.

One asymmetric unit is shown in white, with additional pieces in grey to show alternative shapes. The white pieces together make up the most compact asymmetric unit, which is roughly ellipsoidal in shape with axes 52 × 32 × 26 Å. Later work showed that this asymmetric unit represented a single molecule but, at this stage, we were scrupulous in detailing the alternative interpretations. There is a region of low density that divides the model roughly into two halves, and although we speculated about this we refrained from making any suggestions about its possible significance in our description of the structure (Blake *et al.*, 1962). Instead, we noted that the two halves of the model could be assembled differently, following the crystal symmetry, so as to form a dumb-bell shaped molecule connected at PP′ (Fig. 25.1.2.13*d*).

Our second objective was to determine as far as possible the course of the polypeptide chain and the positions of the disulfide bridges. This proved to be impossible. In comparison with the maps of myoglobin (Kendrew *et al.*, 1958) and haemoglobin (Perutz *et al.*, 1960) at this resolution, it was immediately apparent that this map of lysozyme had a much smaller proportion of clear-cut rod-like features representing α-helices. This was not a surprise, since optical rotatory dispersion measurements (Yang & Doty, 1957) suggested that only 30–40% of the polypeptide chain in lysozyme is in the form of α-helix, as compared with 77% in myoglobin. In addition, Hamaguchi & Imahori (1964) had distinguished the presence of a region of β-sheet in lysozyme before completion of the X-ray analysis. The task of tracing the polypeptide chain, which was difficult with myoglobin, was impossible with lysozyme, since the connectivity of the non-helical regions was often not discernible. The existence of four disulfide bridges, which were expected to have about the same electron density as helices at this resolution, complicated the problem further.

Accordingly, we concluded that defining the shape of the molecule and its tertiary structure would have to await further studies at higher resolution. Meanwhile, Corey and his colleagues (Stanford *et al.*, 1962) and Dickerson *et al.* (1962) published interim accounts of their work at the same time as our work was published (Blake *et al.*, 1962).

At this stage, in the autumn of 1962, RJP left for three months for the MRC Laboratory in Cambridge and then joined Howard Dintzis at Johns Hopkins. At about the same time, DFK joined the team to continue the analysis to high resolution, and LNJ joined DCP as a graduate student and began work related to the activity of the enzyme.

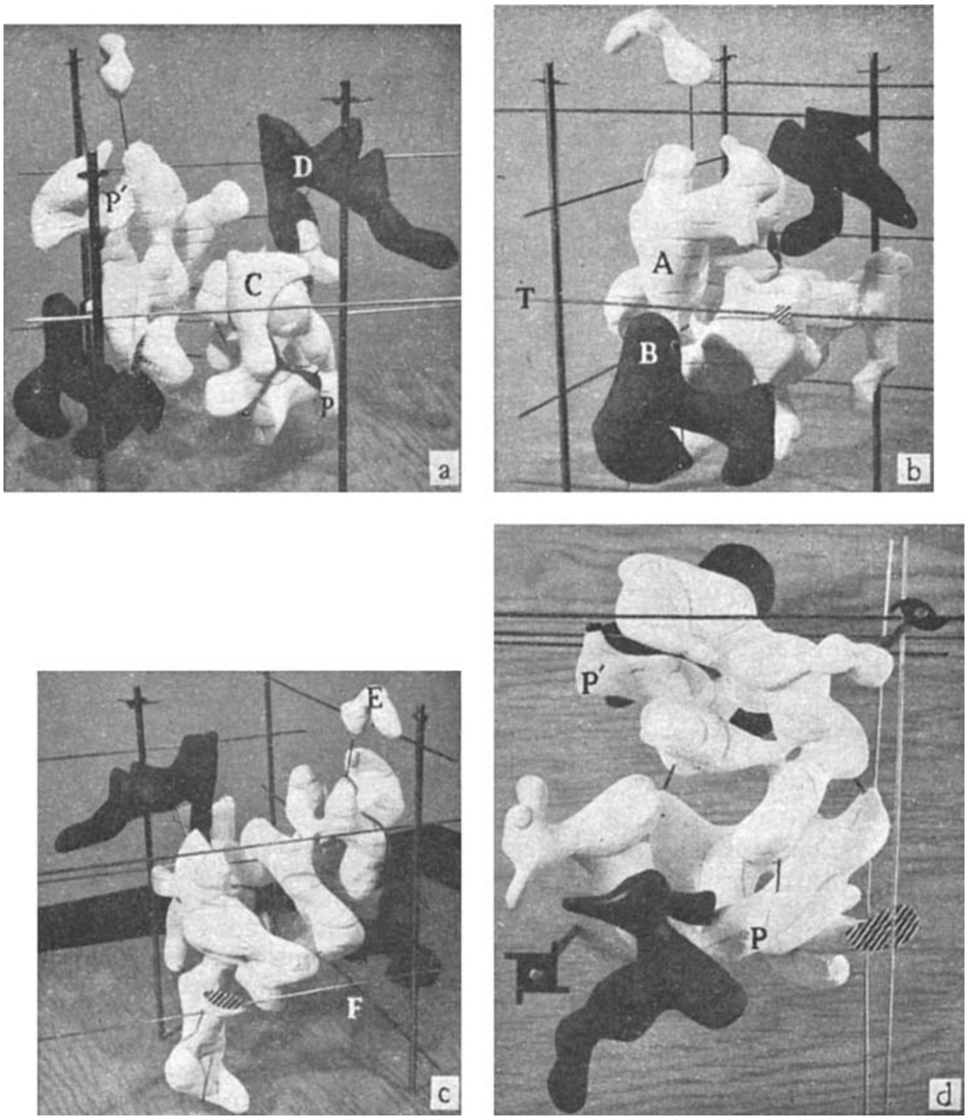


Figure 25.1.2.13
Views of a 6 Å resolution model of the regions in which the electron density exceeds about 0.53 e Å^{-3}. The vertical rods indicate the twofold and fourfold screw axes parallel to **c**. The thinner horizontal rods are twofold rotation axes. At two places, shown hatched, rotation axes pass through continuous regions of density. The white regions alone make up the most compact asymmetric unit of the structure. In part (*b*), the grey section (B) is alternative to the region (A), to which it is related by the twofold axis T. In part (*a*), the grey section (D) is alternative to the white (C), related by the fourfold screw axis. The grey piece (F) in part (*c*) is alternative to the white (E), to which it is related by a unit-cell translation along **c**. The scale is indicated by the framework of symmetry elements, adjacent parallel twofold axes being 18.95 Å apart. Reproduced with permission from *Nature* (Blake *et al.*, 1962). Copyright (1962) Macmillan Magazines Limited.

25.1.3. Analysis of the structure at 2 Å resolution

The structure of chymotrypsinogen (Kraut *et al.*, 1962) at 6 Å resolution was published a few months before the corresponding work on lysozyme. Compared with the work on the globins, however, neither analysis yielded much information on the structures of these proteins, and protein crystallographers generally found them discouraging. Indeed, some went so far as to suggest that only the structures of proteins with a high α-helical content would be amenable to study by X-ray methods. Our reaction, however, as we turned our attention to extending the study of lysozyme to high resolution, was to resolve that each step in the analysis must be conducted as well as possible. We considered carefully, therefore, what improvements might be made to the methods employed hitherto. In particular, we recognized that further work would be needed to identify heavy-atom derivatives suitable for use at high resolution, and we sought improvements in the methods used for data collection, the correction of absorption errors and the use of anomalous scattering in phase determination.

At a purely practical level, one of our major concerns was that our limited access to the London University computer would lead to serious delays in the next stage of the work, which was bound to be even more dependent on computing than the work at 6 Å. Accordingly, we sought support from the Medical Research Council (MRC) for the acquisition of a laboratory-based computer that would be able to handle the computations up to but not including the calculation of a high-resolution electron-density map. Happily, the MRC provided a grant for an Elliott 803B computer, which was installed in the laboratory in March 1963. At the same time, the MRC provided a grant to purchase the commercial version of the linear diffractometer, which was manufactured by Hilger & Watts, Ltd. Since the Elliott 803B had not previously been used for crystallographic computing, this change in our computer involved many members of the laboratory in new programming.

25.1.3.1. Heavy-atom derivatives at 2 Å resolution

The potential usefulness of the three derivatives used at 6 Å resolution for phasing a higher-resolution map was analysed by CCFB and DFK. As the MHTS derivative had much the highest R factor at 6 Å resolution, and the K_2HgI_4 derivative had problems of stability and structure, only the K_2PdCl_4 (PD) derivative seemed likely to be useful for phasing at higher resolutions. An immediate search for additional heavy-atom derivatives was therefore undertaken, which included a re-examination of uranyl nitrate, $UO_2(NO_3)_2$ (UN). Together with other compounds, DFK obtained samples of the then novel $UO_2F_5^{3-}$ ion (UF) from Reuben Leberman at Cambridge, which generated a different pattern of changes in the lysozyme diffraction pattern to any of the previous heavy atoms. When the native phases from the 6 Å map were applied to the UF changes in the centric $hk0$ and $h0l$ zones, they showed a novel two-site binding pattern with a low R factor.

The UF and PD derivatives were examined at 2 Å resolution. Photographs of the centric $hk0$ and $h0l$ zones were taken with a 23° precession angle, the intensities were measured on the Joyce–Loebl densitometer and corrected for Lorentz and polarization effects by a program written by CCFB for the Elliott 803B computer. The heavy-atom parameters obtained in the refinement of these derivatives at 6 Å were used as a starting set for the refinement at higher resolutions. This refinement, like that at 6 Å resolution, was carried out with the program (*Pangloss*) based on Hart's (1961) method, but improved and re-written for the Elliott 803B by DFK.

Initially only the PD and UF derivatives were refined: the mercuri-iodide derivative was not seriously considered to be a potentially useful derivative at high resolution because of the loss of heavy atom with irradiation observed during collection of the 6 Å data. However, it appeared probable that a mean occupancy would be suitable for photographic data, and that the signs predicted by the derivative might be very useful when the other two derivatives gave weak or ambiguous predictions. In fact, the PD and UF derivatives on refinement at 2 Å resolution predicted sets of signs that disagreed to an unacceptable extent, with the PD derivative having a much higher R factor. Refinement of parameters defining the central mercury atom of the mercuri-iodide derivative, followed by calculation of a high-resolution difference-Fourier map using only the most clearly predicted signs, revealed a planar trigonal HgI_3^- ion, whose shape and size were similar to those found by Scouloudi (1965) and Fenn (1964).

Unfortunately, refinement of the mercury and iodine parameters produced a set of protein signs at variance with both previous sets. On the basis of the agreement between F_H(calc) and $F_{PH} - F_P$, both the PD and HgI_3^- derivatives appeared to be non-isomorphous with the native structure. The apparent inability of all three derivatives used at 6 Å to be of use in phasing reflections at higher resolution was disappointing. One possible explanation that was explored for the poor performance of the PD derivative at high resolution, when it showed large changes and a low R factor at 6 Å resolution, was the question of its structure. As with mercuri-iodide, the number of electrons in the halogen substituents exceeded the number in the central palladium atom. Much time was expended in trying to resolve the structure of the square-planar ion without success. This included examining all PdX_4 and PtX_4 complexes, where X is Cl, Br, or I, and calculating double-difference maps between $PdCl_4$ and $PdBr_4$, but without revealing any real indication of structure in the heavy-atom peak.

An extended search for alternative heavy-atom derivatives was therefore begun. Crystals for this purpose were grown as described above (Section 25.1.2.2), except that, on the initiative of DFK, small polyethylene bottles were used instead of glass vials. The polyethylene bottles did not affect crystal size or quality, but allowed crystals to be detached from their surfaces without the damage that occurred when glass vials were used. The search was carried out by the diffusion method. A small number of crystals were isolated together with a known volume of mother liquor, and a solution of the heavy-atom compound, usually at a concentration to make 5 mM in the final solution, was put into a dialysis bag and suspended in the liquid above the crystals. After one or two days, one of the crystals was mounted and an offset 5° screenless precession photograph of the $hk0$ zone was taken. (In these photographs there is no overlap of the zero and first upper levels.) If the approximately 30 unique reflections in this zone showed little or no change in intensities, the crystal was usually not examined further, while those that showed some changes were set aside until a higher-angle precession photograph could be taken. By the use of this technique, a rapid survey of potential derivatives could be made, and those heavy-atom compounds that did not bind to the protein were rapidly dismissed.

Crystals that showed substantial differences in intensity with respect to the native enzyme were used to take $hk0$ and $h0l$ precession photographs. In these studies, the precession angle was reduced to 15° in order to reduce the amount of data to be collected to more manageable proportions, but otherwise the data were collected as for the initial three derivatives. The heavy-atom parameters were also refined in the same way. However, to cope with the disagreements observed between the initial derivatives, a procedure was introduced to combine the refinement of individual derivatives with a refinement of the protein signs in the two centric zones, using all current derivatives together.

When a potential new heavy-atom derivative was identified, the procedure worked as follows. Difference-Fourier maps were calculated, initially at 6 Å and later, when most of the 2 Å signs had been established, at the maximum resolution of the derivative data. When a heavy-atom binding site had been located, its parameters were refined by Hart's method, and the consequences of the refinement, F_P and F_{PH} with their signs and the calculated F_H values, were compared with the corresponding results from other derivatives. The signs of each F_P predicted by the current derivatives were noted, and the most probable sign was determined for each reflection. Initially this was done by inspection, and later by the method of Blow & Crick (1959), with E values provided by the *Pangloss* refinement. The 'best' signs given by this procedure were used to calculate a double-difference map for each derivative with coefficients $[(F_{PH} - F_P) - F_H(\text{calc})]$. When these maps showed either new sites or some new feature at the sites previously included, appropriate alterations were made to the current model of that derivative and further refinement was carried out. If, on the other hand, a derivative showed a high background without interpretable features, it was omitted from further cycles and replaced by another derivative so that a group of four to six derivatives was always in use.

After each derivative had been examined in this way, the set of 'best' signs was updated with the new models of the derivatives, and the procedure was repeated. This procedure worked very efficiently, rapidly indicated a non-isomorphous derivative – or rather one that was significantly less isomorphous than the best – and clearly showed features such as new sites, structure around previously included sites, incorrect or anisotropic temperature factors, and incorrect positional parameters or occupancies. Finally, when the five best derivatives were left in the list, it was immediately apparent, from inspection of the sign predictions in comparison with the E values, which derivatives should be used to phase the high-resolution map.

This method confirmed that MHTS, one of the derivatives used in the 6 Å map but then discarded, was useful at 2 Å resolution and revealed that two new derivatives, UN and UF, were also satisfactory. The poor performance of the MHTS derivative at 6 Å resolution was due to anomalously large changes in the very low resolution reflections, which are sensitive to salt concentrations, but it exhibited good isomorphism at higher resolutions. At about this time, in the summer of 1963, DFK left the laboratory to take up an appointment at Brookhaven National Laboratory.

It became apparent during the course of the heavy-atom search and refinement that the number of suitable derivatives was severely restricted by a feature of the protein. This was the existence of a close pair of very strong binding sites, whose occupation was always accompanied by non-isomorphism. These sites, referred to by the initial derivatives that they bound as the HgI_3^- and PD sites, were found to have a strong affinity for all the complex halogen anions of Pd^{II}, Pt^{II}, Pt^{IV}, Au^{III}, Hg^{II}, Os^{IV}, Ir^{III} and Ir^{IV} that were tried. The sites were mutually exclusive, probably because they shared a protein side chain that acted as one of the important metal-binding groups. The HgI_3 site bound HgI_3^-, $PtCl_4^{2-}$, $PtBr_4^{2-}$, PdI_4^{2-}, $OsCl_6^{2-}$, $IrCl_6^{3-}$ and $AuCl_4^-$. All these compounds caused disordering of the protein structure, as indicated by a decrease in intensities at high resolution, and at least two of the compounds were sensitive to X-irradiation. The PD site also bound $HgCl_2$, $PdBr_4^{2-}$, $PtCl_6^{2-}$, $PtBr_6^{2-}$ and PtI_6^{2-}. All these derivatives were seriously non-isomorphous with the native structure at medium to high resolution, but careful analysis suggested that $PtCl_6^{2-}$ would be useful at low resolution. In contrast, the two uranium compounds, UN [which most probably gave rise to a bound $UO_2(OH)_n$ cation] and UF, gave a substitution pattern entirely different from other derivatives and, in particular, avoided the two sites that gave so much trouble with other complex ions. This finding is in accord with their known tendency to complex with protein oxygens, as opposed to the protein nitrogens that form the binding sites of most other heavy metals.

The two mercury benzene sulfonates that were investigated, MHTS and PCMBS, had a common sulfonate site, but the mercury atoms in the two derivatives were found to be about 3 Å apart. This suggested that they were bound to the protein by their charged sulfonate groups. A similar orientation and position of the benzene ring was implied by the observations that the Hg–SO_3 distance was 5.76 Å in the MHTS derivative, compared with 5.67 Å in the crystal structure of MHTS (Fenn, 1964), and 7.00 Å in the PCMBS derivative, compared with 7.25 Å in the structure of PCMBS. Moreover, the angle Hg(MHTS)–SO_3–Hg(PCMBS) was found in the complexes with lysozyme to be 32°, the same as that calculated from the mercurial structures. This indicates that the benzene sulfonate groups of these two derivatives were fixed in the same position in the lysozyme crystals, and the location of the mercury in the protein crystal depended solely on its position on the benzene ring.

It is interesting to notice that PCMBS, which was investigated first, caused the c axis of lysozyme to lengthen by 1.5% and consequently could not be used at high resolution, while the slightly different MHTS reduced the lengthening to only 0.25% and could be used. This was an early example of an engineered isomorphous derivative. Nevertheless, PCMBS seemed potentially useful at low resolution.

A total of about 50 compounds were used to prepare derivatives (Blake, 1968). Many of these compounds were available commercially but some of the most important, including MHTS and $UO_2F_5^{3-}$, were not. MHTS was designed and synthesised by JWHO, as described above, while the initial sample of $UO_2F_5^{3-}$ was provided by Reuben Leberman, and further supplies were synthesized by CCFB.

In order to prove these new heavy atoms in three dimensions and also to eliminate any 'cyclic' effect due to the use of Fourier methods, 6 Å three-dimensional data were collected early in 1964 for the UF, UN, MHTS, PCMBS and $PtCl_6^{2-}$ derivatives. Two- and three-dimensional Patterson functions were calculated and interpreted afresh. These interpretations were wholly consistent with the results from the high-resolution projections, with the exception of the UN complex. At 3 Å, the $hk0$ and $h0l$ projections showed only two sites, but at 6 Å there appeared to be three additional sites. This point was cleared up when refinement of the 2 Å data collected for phase determination showed that the three extra sites had very high temperature factors, which would result in these peaks being very low at 3 Å resolution. These five compounds were used to phase a second 6 Å electron-density map of the enzyme, as described below, in advance of the analysis at 2 Å, which was based on the only three derivatives that had been found to be reasonably isomorphous at high resolution: MHTS and the two uranyl derivatives, UF and UN.

The discovery of the two uranyl derivatives was not entirely accidental. In the preliminary study at 6 Å resolution, we had been impressed by the potential utility of anomalous scattering in phase determination and had noted the very high value of the imaginary component ($\Delta f'' = 16$) of the anomalous scattering by uranium (Dauben & Templeton, 1955).

25.1.3.2. Intensity measurements

A further advance in diffractometry arose from the observation by DCP that more than one reflection can be measured at the same time in the flat-cone setting of a diffractometer (Phillips, 1964). In the flat-cone setting, the crystal axis is inclined to the X-ray beam as in the equi-inclination setting (Fig. 25.1.2.4*b*), but the motion of the counter is confined to the plane perpendicular to the crystal axis. This flat-cone plane of the reciprocal lattice is midway between the zero-level and equi-inclination levels. When measurements are made in the flat-cone setting from crystals rotating about a reciprocal-lattice axis perpendicular to reciprocal-lattice planes, the crystal and counter settings for reflections in levels adjacent to the flat-cone level are identical to one another and closely similar to those for reflections in the flat cone. This setting is illustrated in Fig. 25.1.3.1.

This property of flat-cone geometry made it possible to modify the linear diffractometer so as to measure three reflections quasi-simultaneously (Arndt *et al.*, 1964). Even more reflections could be measured in this way from crystals with large unit cells but, at

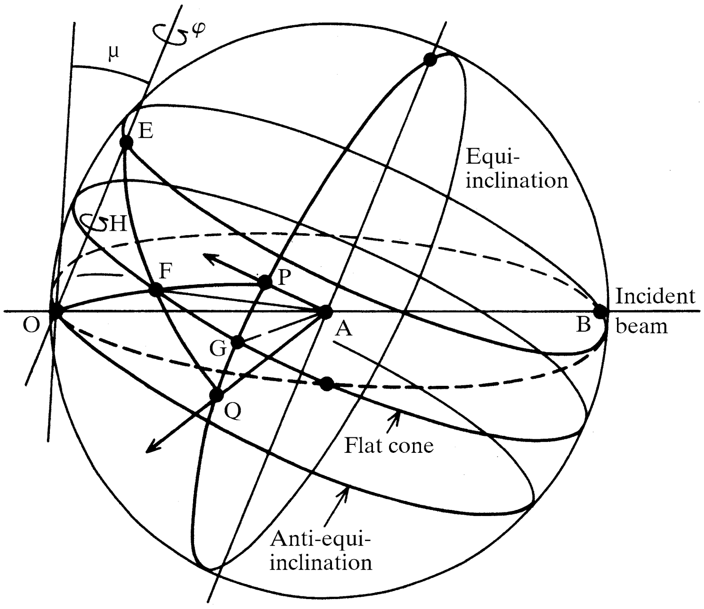


Figure 25.1.3.1
Perspective drawing of the sphere of reflection showing inclination geometry with simultaneous reflections (reciprocal-lattice points P and Q) from levels symmetrically related by the flat-cone setting. Reproduced with permission from Phillips (1964). Copyright (1964) Institute of Physics.

this stage, three 20th Century Electronics side-window proportional counters were mounted in echelon on the counter arm, with appropriate entrance slits and with their windows 0.75 cm apart. The distance of this array from the crystal could be varied between 22 and 50 cm so that the angular separation of adjacent counters lay between 0.034 and 0.015 rad. The separation of reciprocal-lattice levels in which quasi-simultaneous measurements could be made also varied, therefore, from 0.034 to 0.015 reciprocal-lattice units (r.l.u.'s), corresponding to crystal-lattice dimensions of 45 to 100 Å with copper $K\alpha$ radiation.

The array of proportional counters could be set in position on the counter arm by means of horizontal and vertical fine controls. The reflected X-rays passed through adjustable slits before entering the counter windows, and the left and right sides of these slits, or their top and bottom halves, could be blocked to facilitate precise setting.

A disadvantage of using flat-cone as opposed to equi-inclination geometry is that reciprocal-lattice levels measured in a flat cone, and adjacent levels, have a blind region at their centre in which reflections are not accessible. With the linear diffractometer, this blind region falls within the region in which the diffractometer does not automatically set the crystal and counter precisely enough for reliable measurements. The problem is not severe for measurements at low resolution, and it was avoided during the measurement of high-resolution data by the operation of low-angle limit switches that controlled the operation of the scanning and stepping slides. Fortunately, the high symmetry of the lysozyme crystals greatly reduced the seriousness of this problem, since reflections in the blind region close to the rotation axis could usually be measured as symmetry-equivalent reflections in the measurable region.

The first measurements made by this method were 6 Å data for the native protein and the five derivative crystals that had been identified as giving useful phase information at low resolution. Crystals were first mounted to rotate about their *a* axes, and the crystal-to-counter distance was 38.5 cm. Unfortunately, this long crystal-to-counter distance gave rise to weak measured intensities, because of the significant absorption of the reflected X-rays by the air in the counter arm. For this reason we considered filling the counter arm with hydrogen, but came to the conclusion that a simpler approach would be to make the bulk of the measurements from crystals mounted to rotate about their $[\bar{1}10]$ axes.

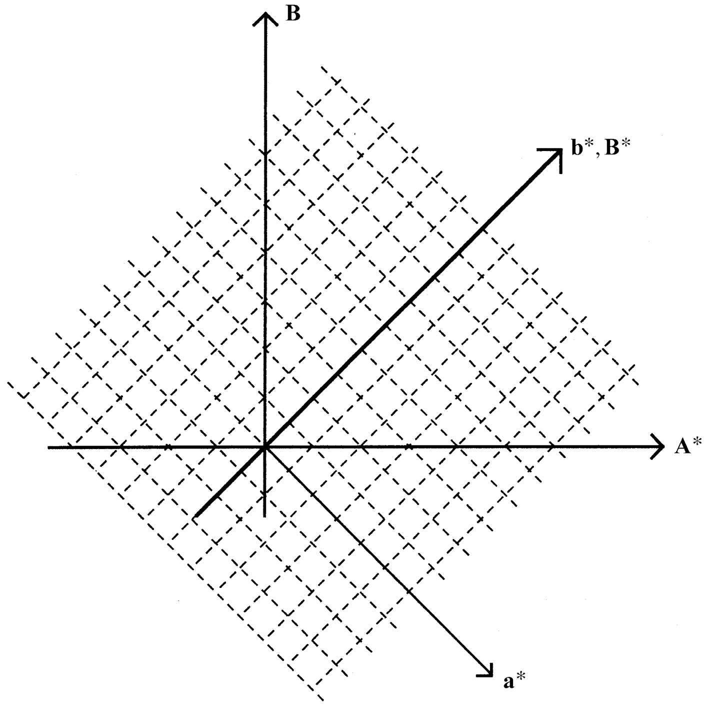


Figure 25.1.3.2
The a^*b^* plane of the reciprocal lattice oriented to rotate about the $[\bar{1}10]$ axis. The indexing of reflections in a monoclinic unit cell with axes A^* and B^* is also shown. A^* coincides with [110] in the tetragonal lattice and B^* coincides with b^*. The rotation axis is the B axis in the monoclinic cell. The reciprocal-lattice dimensions with Cu $K\alpha$ radiation are $a^* = b^* = 0.0195$; $c^* = 0.0406$ r.l.u., while the cell diagonal in the a^*b^* plane is 0.0276 r.l.u.

In this setting, the pyramidal end of the crystals fitted against the wall of the capillary tube in which they were mounted (Fig. 25.1.2.5*b*). The roughly square cross section of the crystals perpendicular to the rotation axis tended to minimize the absorption variation, which was significant with crystals of linear dimensions of about 0.5 mm. The use of large crystals was necessary with a relatively weak X-ray source running at 800 W and with a foreshortened focal spot of 0.4 × 0.4 mm.

It was convenient to index the reflections in a monoclinic cell, as shown in Fig. 25.1.3.2. In this orientation, the reciprocal lattice presented a diamond pattern to the triple-counter array, whose windows were set parallel to the rotation axis. The reflections therefore occurred in two intersecting sets of levels, the odd and even levels, which had to be measured separately. However, the need to collect alternate levels in this way conferred the advantage that the counters could be positioned closer to the crystal (27.1 cm) than in the *a*-axis mounting so that >70% of the reflected X-rays were transmitted to the counters. Despite the complexity that this geometry introduced at the data processing and reduction stages, the significant advantages that it offered at the experimental stage ensured its use. The use of a c^* mounting would have been even more advantageous, but it was ruled out both by the difficulty of mounting the crystals in this orientation and by the fact that the counter arm could not be set to the required length of 18.5 cm.

Native data were collected both from crystals rotated about the *a* axis and from crystals rotated about $[\bar{1}10]$ so that they could be scaled together to form a consistent set of three-dimensional data. For the derivatives, however, data were collected only from crystals rotating about their $[\bar{1}10]$ axes and these data were scaled together to form two sets, one comprising levels with odd indices corresponding to the rotation axis and the other with even indices. These measurements, comprising some 1200 reflections, could be made from a single crystal exposed to the X-ray beam for about 20 h. The odd and even data sets for the derivatives were then scaled separately to the native data to give complete sets of three-dimensional data. These low-resolution measurements were made during the first half of 1964 and, after processing by the methods described below, they were used in September 1964 to calculate a new image of the structure at 6 Å resolution.

25.1.3.3. The second low-resolution map at 6 Å

Our purpose in calculating a new electron-density map at 6 Å was fourfold. First to ascertain whether the procedures used to identify the five derivatives thought to be satisfactory at this level of resolution had worked satisfactorily. Second, to judge the quality of the measurements made by the triple-counter diffractometer. Third, to explore the effects of the modified method of applying absorption corrections to the intensities that are described below, although these were not expected to have a very great effect at low resolution. Fourth, to examine the effectiveness of the new procedure for incorporating anomalous-scattering information in the phase determination, which is also described below.

Comparison of the two sets of structure amplitudes gave a conventional *R* value of 0.075, which is not particularly good – perhaps because of the comparatively large background values associated with these low-angle measurements. However, the mean figure-of-merit obtained in the new phase calculations was 0.97 as compared with the 0.86 obtained originally. The root-mean-square difference in electron density between the two maps was 0.012 e Å^{-3}, from which it may be judged that the two maps were very similar. Nevertheless, the outline of the molecule was certainly clearer in the new map, and within the molecule there was improved continuity, suggesting the course of a folded polypeptide chain, and there were a number of stronger rod-like features suggestive of α-helices. Two of these were prominent, running upwards from right to left, in the view of the new model shown in Fig. 25.1.3.3.

The result was very encouraging, and we therefore went ahead immediately with data collection at 2 Å resolution, using essentially the same methods. At the same time, we began to plan low-resolution studies of inhibitor binding to lysozyme, from which we hoped to derive information about the nature of the enzyme–substrate complex.

25.1.3.4. Intensity measurements at high resolution

At high resolution, measurements were made from only six levels from native crystals set to rotate about the *a* axis while a complete set of data was collected from crystals rotating about $[\bar{1}11]$. As at low resolution, the *a*-axis levels were important for scaling together the odd and even sets of levels measured from crystals rotated about the tetragonal $[\bar{1}11]$ axis. No high-resolution measurements were made from derivative crystals rotated about the *a* axis, the intention being to scale the odd and even subsets of derivative data directly to the native.

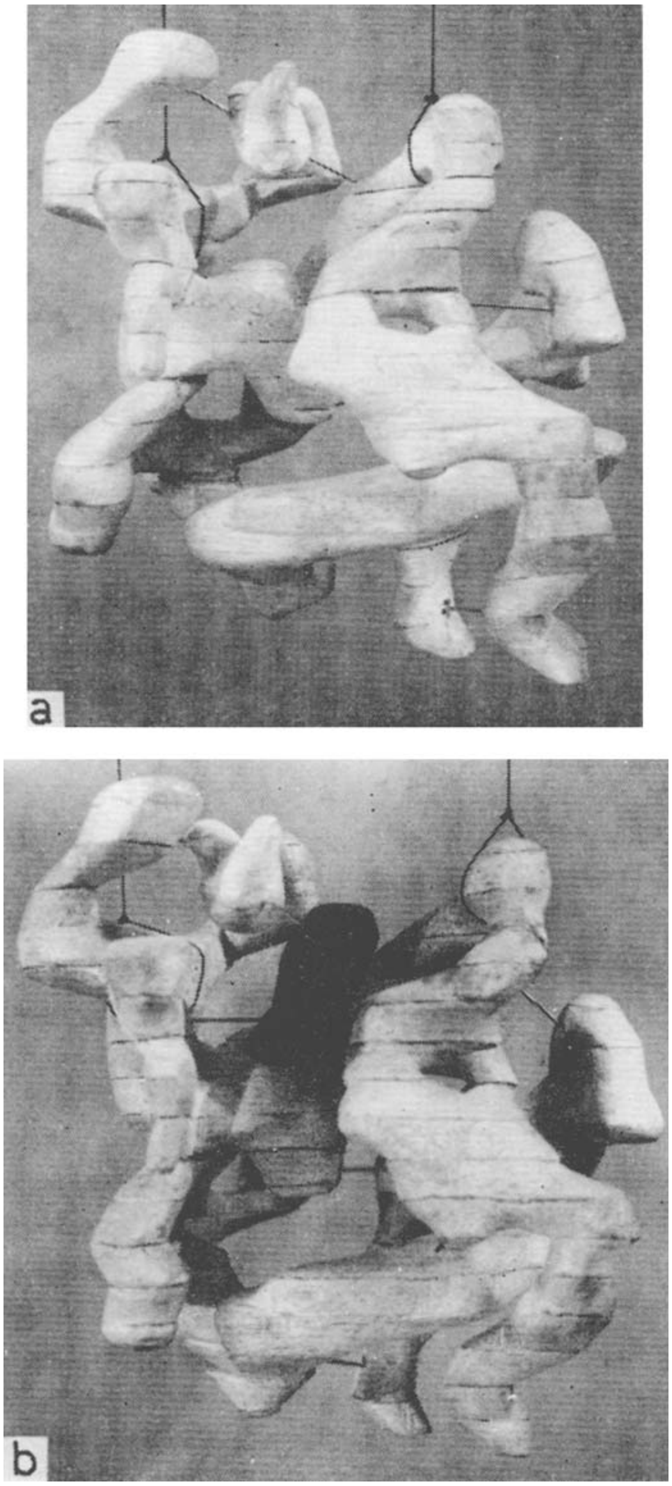


Figure 25.1.3.3
Solid model of the electron density greater than about 0.5 e Å^{-3} in the second study of lysozyme at 6 Å resolution. This view of the model is equivalent to a view of the original model seen horizontally from the right of Fig. 25.1.2.13(*c*). (*a*) The new model has a marked cleft running roughly vertically down the other side of the model, corresponding to the one that can be seen in Fig. 25.1.2.13(*c*). (*b*) The cleft was shown to bind inhibitor molecules. The black density is that observed for the lysozyme–GlcNAc complex at 6 Å resolution.

25.1.3.4.1. Experimental methods

The crystals can be thought of as roughly flattened on $(00\bar{1})$ with liquid between $(00\bar{1})$ and the capillary tube in which they were mounted. In order to minimize the effect of the liquid on absorption, we therefore measured the reciprocal-lattice hemisphere with *C* and c^* positive. Reflections were scanned along *A*, and the origin of the scanning slide was offset for upper levels so that, for example, the centre of the scanning slide was the $H = -4$ for the level $K = 8$.

Crystals could be set fairly well by eye. Viewed through the diffractometer microscope, they looked roughly square in the [001] direction, and the arcs of the goniometer head were set so that the edges were horizontal and vertical. When the crystal was turned through 90° from this position, the reflection of a light held level with the microscope could often be seen in the true (110) face. Setting the crystals on the goniometer head with [001] parallel to one of the arcs facilitated subsequent adjustments.

Fine adjustment of crystal orientation was achieved by setting on the $\bar{4}40$ reflection. For this purpose, C – the vertical slide – was set to 0.1104 and the inclination angle to $\mu = 3° \, 5'$. Then, if one arc was set fairly well by eye, the other could be oriented parallel to the incident X-ray beam and adjusted to locate the reflection. The two arcs were then adjusted in turn to give the optimum setting, in which the crystal was rotated about the normal to the ($\bar{4}40$) plane.

The orientation in the AC reciprocal-lattice plane was determined by returning the vertical slide to zero and setting the upper (scanning) slide to 0.1104 with the lower (stepping) slide at zero. The crystal was then rotated until the monoclinic 400 reflection appeared. At this stage, checks were made with the top/bottom and left/right slits to make sure that the crystal was well centred in the X-ray beam and that the counter apertures were well positioned. Similar checks were made with the crystal rotated about 180° to the monoclinic $\bar{4}00$ reflection.

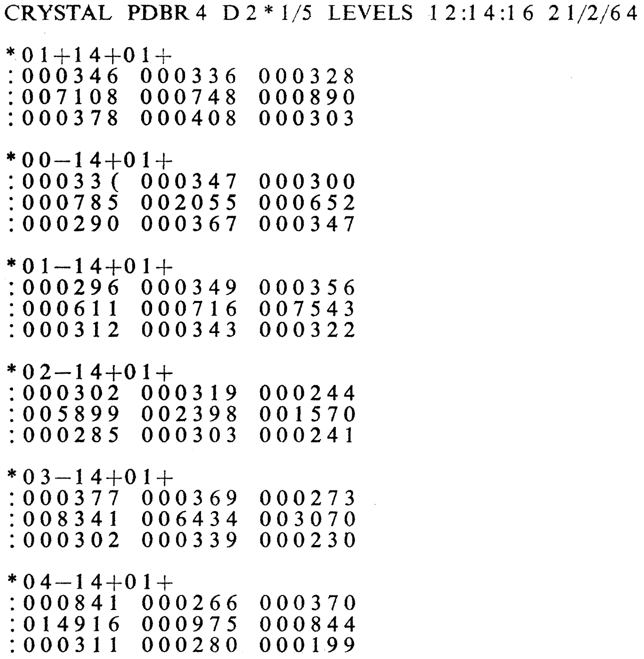
The final check on crystal orientation was to locate the 008 reflection near 0.3248 on the lower (stepping) slide, with the other slides set at zero. There was some variation in the value of c^* for different crystals, and c^* was often closer to 0.0404 than to 0.0406 r.l.u.

At this stage, the first measurements were made of the intensities of the reflections (monoclinic) 400, 16,0,0 and 008. These reference reflections were remeasured at intervals during the measurement of each triplet of reciprocal-lattice levels as a check on the stability of the whole system and irradiation damage to the crystal. The measurements were manually entered on the diffractometer output tape and monitored by the data-processing program.

In order to set the diffractometer for a particular triplet of levels, we found it convenient to leave the lower slide set for the 008 reflection, to adjust the vertical slide and the tilt angle (μ) for the levels in question, and then to run out along the stepping slide until a suitable high-angle reflection was found. The crystal rotation angle was then optimized for this reflection.

Finally, careful checks were made to ensure that the reflections to be measured fell in the reciprocal-lattice hemisphere with L positive, and that they were indexed in a right-handed axial system. The automatic run was then begun at the 2 Å limit on the stepping (C^*) slide, and all reflections were scanned from $H_{\max}$ to $-H_{\max}$ on the scanning slide. Virtually all reflections within (and in some directions a little beyond) the 2 Å limit were measured in this way in fourteen triplet levels, seven even and seven odd. About 2000 reflections were measured in a typical overnight run, and data collection for each species of crystal took slightly more than two weeks.

The derivative crystals required for these measurements were prepared by the diffusion method described above, and about 20 of them were mounted at the same time at the beginning of the data-collection process to ensure that they were all the same. At the end of a run, careful measurements were made of the peak intensity of the $\bar{4}40$ reflection as the crystal was rotated through steps of 15° about the crystal-mounting axis (φ). These measurements were used during data processing in an improved method of absorption correction devised by North *et al.* (1968).



```
CRYSTAL PDBR4 D2*1/5 LEVELS 12:14:16 21/2/64

*01+14+01+
:000346 000336 000328
:007108 000748 000890
:000378 000408 000303

*00-14+01+
:00033( 000347 000300
:000785 002055 000652
:000290 000367 000347

*01-14+01+
:000296 000349 000356
:000611 000716 007543
:000312 000343 000322

*02-14+01+
:000302 000319 000244
:005899 002398 001570
:000285 000303 000241

*03-14+01+
:000377 000369 000273
:008341 006434 003070
:000302 000339 000230

*04-14+01+
:000841 000266 000370
:014916 000975 000844
:000311 000280 000199
```

Figure 25.1.3.4

Paper-tape output from the triple-counter linear diffractometer, showing the indices of the central reflection and the background, peak, background counts for the three reflections. Reproduced with permission from Arndt *et al.* (1964). Copyright (1964) Institute of Physics.

25.1.3.4.2. *Diffractometer output*

The output from the triple-counter diffractometer again consisted of a plain-language output for immediate checking of the results and output at this stage on eight-hole-punched paper tape for immediate transfer to the computer. An example of this record is shown in Fig. 25.1.3.4.

25.1.3.5. *Data processing*

Although the in-house Elliott 803B computer provided much improved facilities over those previously available, its limited store capacity (4096 words of magnetic core memory, but no magnetic drum, disk, or tape facilities) resulted in data processing being carried out in a series of stages, each of limited complexity. Input and output was by means of eight-hole paper tape, which was used as the medium for intermediate storage. The output tape from each stage was used for input to the next stage, together with appropriate parameters. Manual input was restricted to such parameters as unit-cell dimensions and ordinates of the absorption curves, with all other input being in the form of computer-generated output from a previous stage, starting with the diffractometer output tapes. This approach resulted in a great reduction in manual labour and intervention compared with the low-resolution work on lysozyme and the high-resolution stage of myoglobin, and it was probably the first crystal-structure determination that was fully computerized.

As for the initial low-resolution work, the first stage of data processing was input and checking of the diffractometer output, with the programs modified appropriately to deal with the sets of three reflections measured simultaneously. The background-corrected intensities were then output together with a table of the reference reflections used to estimate the extent of any radiation damage. Corrections were then applied for radiation damage if required, followed by application of Lorentz–polarization corrections, followed by absorption corrections.

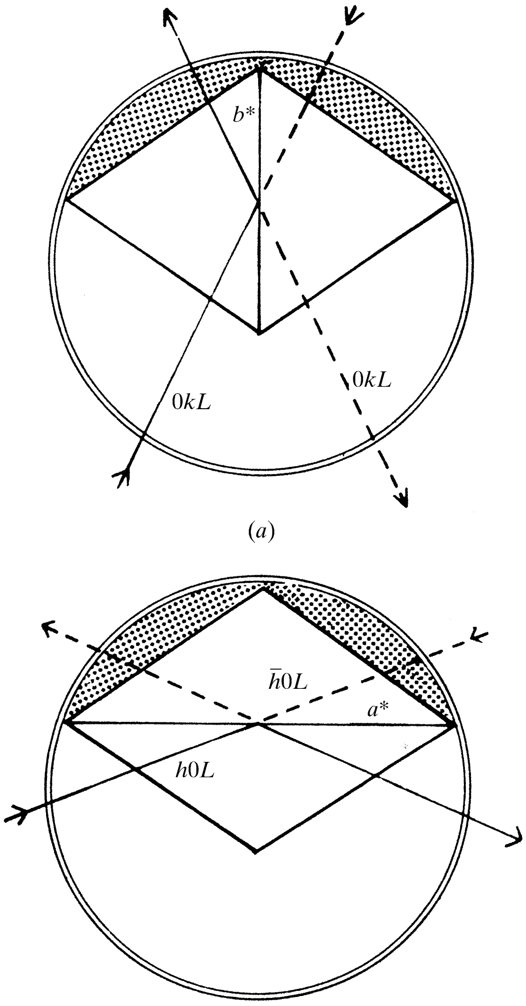
25.1.3.5.1. Absorption corrections

Although absorption of X-rays by protein crystals is low compared with crystals having a preponderance of heavier atoms, corrections for absorption are required in order to give F values that are sufficiently precise for calculation of the relatively small changes due to the introduction of heavy atoms or anomalous dispersion. The mounting of protein crystals within a glass capillary, normally with a small amount of mother liquor between the crystal and the capillary wall, presents a complicated situation for absorption calculations. Although Wells (1960) wrote a computer program to deal with the situation, a severe impediment to the use of theoretical methods of correcting for absorption results from the very great difficulty in obtaining precise measurements of the mounted crystal, the liquid meniscus and the capillary-tube walls. An alternative approach was to use a semi-empirical method and, for low-resolution lysozyme studies, Furnas' (1957) method had been employed, as described above.

Despite the fact that the method of Furnas had been successful in improving the agreement between symmetry-related reflections in the earlier studies, the implicit assumption that the absorption depended upon the mean direction of the incident and reflected X-ray beams became clearly less valid as the Bragg angle increased. We therefore implemented a development of the method in which the absorption correction applied to any reflection was given by the mean of the two values for the directions of the incident and reflected beams (North *et al.*, 1968).

Although the method was easy to apply and was of significant value, as judged by the improved agreement between symmetry-related reflections, it nevertheless provided only a partial correction for absorption, because of the assumption that absorption is dependent solely on the directions of the incident and reflected beams. The limitations of this assumption are particularly important where precise values are required for Friedel pairs of reflections in order to make use of anomalous-scattering differences in phase determination. Fig. 25.1.3.5 shows two contrasting situations that can arise when the environment of a crystal is asymmetrical because of its mounting. The Friedel pair of reflections shown in Fig. 25.1.3.5(*a*) would suffer similar absorptions, whereas the pair shown in Fig. 25.1.3.5(*b*) would have significant differences because of the location of the mother liquor. In the 2 Å structure determination of lysozyme, an approximate correction was made for this effect. From Fig. 25.1.3.5, it is clear that the absorption error arising from the asymmetric distribution of the mother liquor is 0 for reflections with $h = 0$ and becomes increasingly great as h increases. The assumption was made, therefore, that the required correction was a function only of h and the reflections hkL with constant L were divided into groups with constant h and $-h$. The ratios $\sum_k I(h)/\sum_k I(-h)$ were then plotted against h, as shown in Fig. 25.1.3.6.

Such plots were frequently found to be linear, and the corresponding linear correction was then applied to each row on the more highly absorbed side in order to bring its mean intensity up to that of the other. Where the plots were not linear, so that a simple form of correction was not applicable, the entire set of measurements was usually rejected.



(*a*)

(*b*)

Figure 25.1.3.5
Asymmetric mounting of a protein crystal with its mother liquor in a capillary tube. Anomalous-scattering differences would be seriously affected in (*b*) but not in (*a*). Reproduced with permission from North *et al.* (1968).

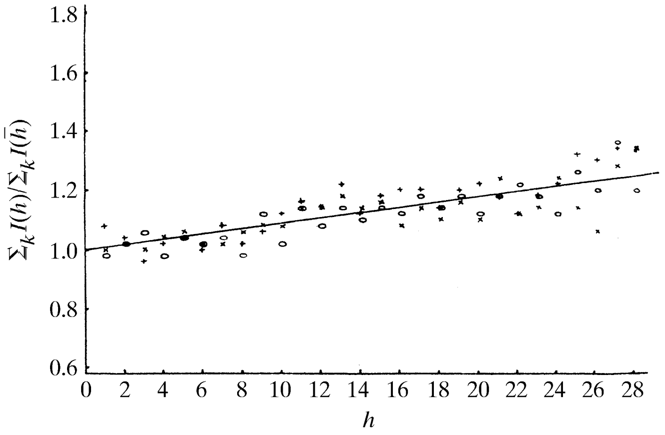


Figure 25.1.3.6
Plot of the ratio $\sum_h I(h)/\sum_h I(-h)$ against h. In this example, a linear correction may be safely applied to equalize the average intensities in opposite rows (North *et al.*, 1968).

25.1.3.6. Further stages of data processing

Many of the programs for subsequent stages of data processing were written by VRS, who had discussed the work with DCP at a meeting in Madras in January 1963 and who joined the team in October 1963.

As described in Section 25.1.3.2, the native data comprised three sets, one of six reciprocal-lattice levels measured from crystals rotated about the *a* axis, and the other two comprising the odd and even subsets of levels collected about the $[\bar{1}11]$ axes. Within these three sets, the odd and even $[\bar{1}11]$ levels had no rows in common with one another, but each had rows in common with the *a*-axis data. Extraction of the related rows permitted the calculation of scale factors by the method of Hamilton, Rollett & Sparks (Hamilton *et al.*, 1965). Application of these scale factors produced a complete set of self-consistent native data.

The derivative data, which consisted of the odd and even levels collected about the $[\bar{1}11]$ crystal axes, were scaled to the native data by the application of two scale factors derived by comparing the totals of the intensities in the odd and even subsets. Because of the high symmetry of the tetragonal space group, up to eight measurements were available for many symmetry-equivalent reflections; for the heavy-atom derivative crystals, these formed four sets of Friedel-related pairs. At this stage, the X-ray amplitude data were on 40 paper tapes for each of the native and three derivative crystals, one for each reciprocal-lattice level. The native and derivative data were then compared level by level, but it was found that the ratio was very nearly a constant, independent of $\sin\theta$. A further numerical factor was applied to all data to bring them approximately to an absolute scale. In the final stage of data processing, the data from the native and the three derivative sets were brought together into a single list containing the native *F*'s and the Friedel pairs for each of the three derivatives. In addition, all of the centric reflections were extracted in order to provide the data for refinement of the heavy-atom parameters.

At this late stage in the data processing, we noticed that some levels agreed significantly less well than the norm with intersecting levels. On 23 October 1963, we had to recognize that the high-resolution measurements had been made from two different types of crystals with essentially identical unit-cell dimensions, but with subtly different diffraction patterns. We designated these two crystal types type I and type II, and set about the task of producing sets of data of one type to use in the structure analysis.

25.1.3.7. The crystal-type problem

This discovery was particularly galling because, although at the time variations in diffraction patterns had been reported for some protein crystals, we had failed to notice that this phenomenon had been mentioned years earlier in a study of lysozyme (Corey *et al.*, 1952).

A preliminary analysis of the differences in the diffraction patterns suggested two important characteristics:

(1) the differences tended to increase with resolution and
(2) the differences appeared to be more consistent with two discrete diffraction patterns than a continuum of patterns lying between two extremes.

The two diffraction patterns were characterized operationally by specific patterns of intensities in the 3–4 Å resolution range (where the differences appeared to be maximal), and particularly by a few adjacent pairs of reflections whose relative intensities interchanged in the two types. The principal diagnostic reflections were 11,11,4 and 11,11,5. In data associated with crystal type I, $I(11,11,4) > I(11,11,5)$, while in crystal type II, $I(11,11,4) < I(11,11,5)$. The 11,11,*L* rows of reflections from the two types of crystals had the structure amplitudes shown in Table 25.1.3.1.

Table 25.1.3.1

Structure amplitudes of the 11,11,*L* reflections from crystal types I and II

Crystal type	Reflection					
	11,11,0	11,11,1	11,11,2	11,11,3	11,11,4	11,11,5
I	29.2	21.8	10.6	18.2	23.1	17.0
II	36.9	25.7	19.7	27.7	26.4	36.8

In order to stand any chance of successfully calculating the lysozyme map by isomorphous replacement, we had to sort all the data so far collected, both native and derivative, into the two types and then recollect 'rogue' data sets in order to assemble complete data sets of one particular type. The alternative was to recollect the whole data on a sounder basis, which we were loath to do, especially as other teams seemed likely to be well advanced in their solution of the lysozyme structure.

It appeared that the bulk of the data that had already been collected using the diffractometer was what we had called type II. We also observed that nearly all the photographic data collected in the heavy-atom proving stage was of type I. This observation was of great importance because it gave us a sound basis for defining the differences between the type I and II diffraction patterns, and it also provided a vital clue in identifying from their shapes the crystals that gave the two types of diffraction pattern. This was very important to us in the selection of crystals to replace the rogue data sets. The crystals that gave the best results for the photographic work tended to be relatively small and flattened along the tetragonal fourfold axis direction, while those that gave the best diffractometer data were the larger isometric crystals, which were more extended along the crystal fourfold axis (Fig. 25.1.2.1). These crystals could be definitely associated with the type I and type II diffraction patterns, respectively. Batches of lysozyme crystals grown according to the procedures defined earlier usually contained both types of crystal. This suggested that the two crystal forms might have originated when the pH of crystallization was on the borderline between two crystal forms of lysozyme. This hypothesis was supported by the observation that crystals grown at a somewhat higher pH had diffraction patterns closely similar if not identical to type I.

The more thorough analysis of the differences between the diffraction patterns for types I and II that these findings permitted showed that the differences in the intensities of equivalent reflections were resolution-dependent. They were very small in the 6 Å region (which probably accounts for the differences not being observed earlier), increased to a maximum at the position of the normal 4 Å peak in the protein diffraction pattern, and fell off at higher resolutions. This pattern is consistent with a lack of isomorphism between the two types of crystal of the kind that may be caused by a slightly different orientation of the lysozyme molecule in the unit cells of the two crystals (Crick & Magdoff, 1956). Such effects may be brought about, for example, by slightly different charge distributions in the protein molecules, rather than by the presence of additional diffracting material in one crystal type or the other. [This conclusion was later confirmed by a detailed analysis of the two structures by Helen Handoll (1985).]

These observations suggested it was safe to go ahead with trying to determine the structure of lysozyme in either type of crystal, and the decision to proceed with the type II crystals was solely on the basis that the bulk of the data already collected were of this type. Knowing the characteristic pairs of reflections that distinguished the two crystal types, we found it relatively straightforward to ascribe each data set to a particular crystal type, even for the heavy-atom derivatives, because the type differences were much larger than the heavy-atom changes at

Table 25.1.3.2
Heavy-atom parameters for the 2 Å structure

O is the occupancy of the heavy-atom site (electrons); B is the isotropic temperature-factor constant; E is the root-mean-square difference between observed and calculated heavy-atom differences for centric reflections (electrons); N is the number of centric reflections in the range $0.01 < \sin^2\theta < 0.15$ used in the refinement and R is the reliability index for observed and calculated heavy-atom changes of centric reflections.

Derivative	Site	x	y	z	O	B (Å^2)	E	N	R
MHTS	I	0.2068	0.6138	0.0507	39.2	17.8	58	1247	0.60
	II	0.2415	0.6393	0.9326	8.8	14.9			
UF	I	0.1783	0.5849	0.7204	55.5	21.0	74	1277	0.52
	II	0.0974	0.8976	0.4650	29.3	24.3			
UN	I	0.0961	0.8938	0.2664	47.1	19.2	80	1140	0.57
	II	0.1898	0.5901	0.7168	42.1	124.8			
	III	0.0446	0.7266	0.5150	9.0	190.2			
	IV	0.0869	0.8976	0.4866	11.4	68.4			
	V	0.2024	0.6388	0.6781	28.6	42.8			

approximately 4 Å resolution. All triplet levels of data that belonged to the type I diffraction pattern were extracted from the total data sets and replaced by equivalent data sets recollected from confirmed type II crystals. This process, whose success was carefully tested and confirmed during the data processing and reduction stage, resulted in consistent data sets at 2 Å resolution for the native lysozyme and all three isomorphous derivatives that were derived from type II crystals.

25.1.3.8. *Final refinement of heavy-atom parameters*

At this stage, we were able to review the refinement of the heavy-atom parameters in relation to the type II crystal data that were now available. Prompted in part by the observation that the heavy-atom positions in myoglobin appeared in regions of high negative density in the final electron-density map, we considered the need to adopt an improved way of modelling the heavy atoms, particularly for the complex uranium compounds, in which the uranium atoms were coordinated to several oxygen or nitrogen atoms. Calculation of the R factors and occupancies of the heavy atoms as a function of $\sin\theta$ showed that the heavy-atom compounds, particularly the uranyl complexes, could not be modelled well by a single atom, but that it was necessary to take into account both the O and the N atoms in the complex and the water that had been displaced by the heavy-atom cluster. The difference electron-density maps did not show the orientation of the O and N atoms, and our first thought was to model the complex in terms of a central U atom, surrounded by a spherical shell of electron density representing the O and N atoms minus a sphere of electron density representing the displaced water. Trial calculations suggested that this model would give improved agreement between observed and calculated $|\Delta F|$ values, most significantly for the uranyl derivatives. On second thought, it occurred to us that a more satisfactory approach might be to derive empirical scattering-factor curves by fitting the curves representing the variation of heavy-atom occupancies with $\sin\theta$ to polynomial functions of $\sin^2\theta$. This was easily accomplished by the use of a standard curve-fitting program and, for ease of use in the phase program, these scattering curves were fitted to curves of the type $f(H) = a + b\sin^2\theta + c\sin^3\theta + d\sin^4\theta$. With this modification, *Pangloss* was used with the type II crystal centric data to obtain the heavy-atom parameters shown in Table 25.1.3.2.

The temperature factors shown in Table 25.1.3.2 most deserve comment. All those obtained for the significant sites (of which there are no more than two for any derivative) are comparable with the overall value for the protein crystals themselves. The very large values obtained for other sites show that these sites are of little importance at high angles and may not represent real sites of heavy-atom attachment. The minor site of MHTS is clearly the SO_3^- group of this molecule (see Section 25.1.3.1).

25.1.3.9. *Calculation of phase values*

Blow (1958), in his determination of the phase angles of the noncentrosymmetric [100] zone of horse haemoglobin, and, later, Cullis *et al.* (1961, 1962), in their determination of the three-dimensional structure of horse haemoglobin, used anomalous-scattering data to supplement the information available from the isomorphous-replacement differences. In each of these studies, phase determination had been carried out by constructing probability curves from the multiple-isomorphous-replacement data and, when the most probable phase angle had been deduced, the anomalous-scattering data were examined. For many reflections, they allowed a choice to be made between two apparently equally probable values of phase angle given by the isomorphous data. This procedure was clearly rather arbitrary and subjective, and a method of combining anomalous scattering with isomorphous replacement in a more rigorous way was described by Blow & Rossmann (1961). In their method, which was subsequently employed for the low-resolution work on lysozyme, use was made of the fact that the mirror image of the Argand diagram for a $\bar{h}\bar{k}\bar{l}$ reflection is similar to the Argand diagram for the hkl reflection, but for the reversal of the sense of the imaginary part of the heavy-atom contribution. The data for the $\bar{h}\bar{k}\bar{l}$ reflections may therefore be treated as though they came from a separate isomorphous compound, with parameters identical to those of the original compound, but with the opposite sign for the imaginary component of the atomic scattering factor.

In the low-resolution lysozyme phase determination (Section 25.1.2.7), intensities of the Friedel pairs of reflections were measured for each of the three heavy-atom compounds, and the problem was treated as if there had been a total of six heavy-atom compounds. Although the method had been found helpful to some extent, analysis of the phases showed that the anomalous-scattering data had played comparatively little part in determining the positions of the centroids of the phase probability distributions, even for reflections with apparently significant anomalous differences.

ACTN observed that this apparent contradiction is because of the fact that the anomalous differences between Bijvoet pairs of reflections measured from the same crystal are inherently more accurate than the isomorphous differences that are measured from different crystals and subject to different systematic errors (North, 1965; Phillips, 1966). Indeed, analysis of the equivalent reflections from native and derivative crystals (Section 25.1.3.6)

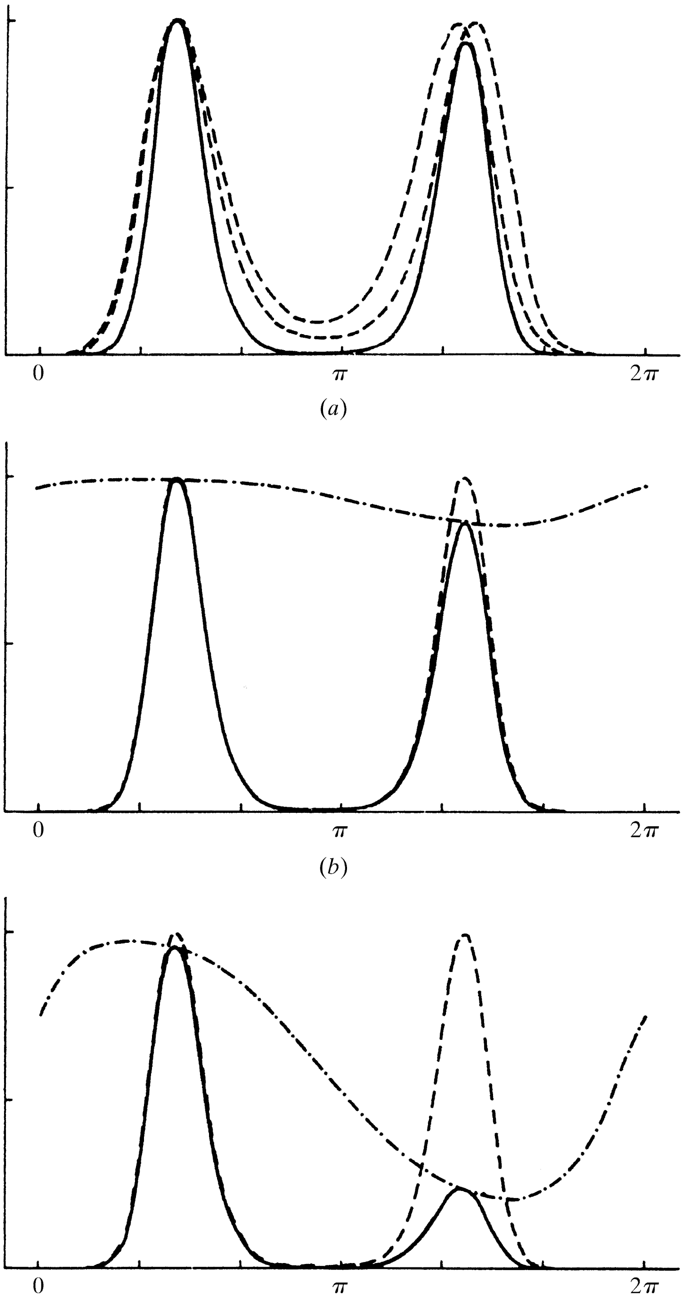


Figure 25.1.3.7
(*a*) Phase probability curve for a Bijvoet pair of reflections (broken lines) with the joint probability curve (full line) derived by the method of Blow & Rossmann (1961). (*b*) Isomorphous-replacement phase probability curve derived from the mean of F_{PH^+} and F_{PH^-} (broken line); anomalous-scattering probability curve (chain line); and joint probability curve (full line) derived by the method of North (1965), using E'(anomalous) = E(isomorphous). (*c*) As (*b*), but with $E' = (1/3)E$. Reproduced with permission from North (1965).

showed that the r.m.s. error E', corresponding to the anomalous differences, was about one-third of E, the error in the isomorphous differences. The result of incorporating this distinction in the phase program is illustrated in Fig. 25.1.3.7. Phase calculations for the new 6 Å and 2 Å maps of lysozyme were therefore carried out by using ACTN's method, with E' set at one third of E.

The data tapes containing the F values for the native and the Friedel pairs of F values for the three derivatives were used as input to a phase program written by ACTN. For acentric reflections, phase probabilities were calculated as described in the previous section, and the centroids of the distributions were determined in order to derive a 'figure of merit', which was applied to the structure amplitudes, as first proposed by Blow & Crick (1959), so as to produce a 'best' Fourier map. For the quite high proportion of centric reflections in the lysozyme diffraction pattern, phase probabilities were calculated by the formula appropriate to the case in which the native and derivative F's are collinear with each other and with the vector due to the heavy atom.

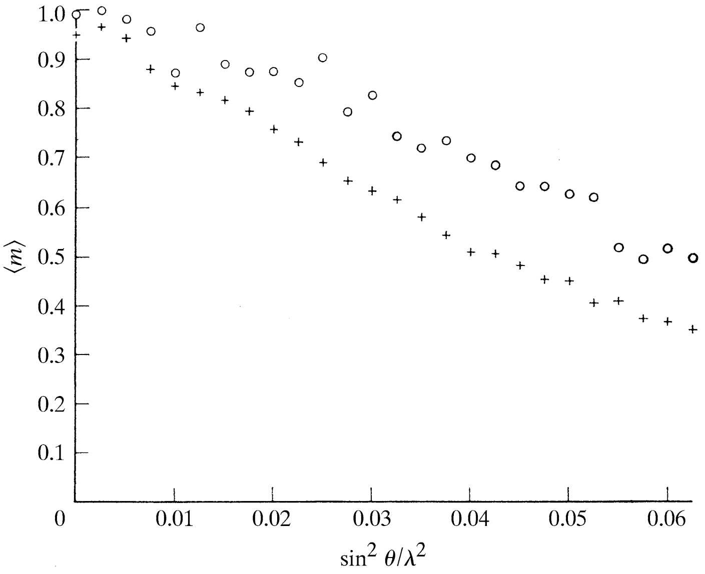


Figure 25.1.3.8
Variation of the mean figure of $\langle m \rangle$ with $\sin^2 \theta/\lambda^2$ (crosses represent acentric reflections, open circles represent centric reflections).

The phases of the 9040 reflections were calculated on the Elliott 803B computer and had a mean figure of merit of 0.60. The variation with angle was very similar to that obtained with sperm-whale myoglobin and is shown for the centric and acentric reflections separately in Fig. 25.1.3.8.

25.1.3.10. The electron-density map at 2 Å resolution

GAM joined the team in early 1964 with the specific task of writing a program for calculating the 2 Å electron-density map. Fortunately, at this stage, the University of London Computing Centre was in the course of acquiring a Ferranti ATLAS computer, which was then one of the most powerful computers available for this kind of work, and GAM eagerly set about taking advantage of its power. Whereas we ourselves had operated the previous MERCURY computer at night, there were machine operators for the ATLAS with whom we left our input tapes; these were very large reels of paper tape comprising all of the amplitude and phase data for the 2 Å set. Following two or three unsuccessful attempts to load the whole of the data without tearing the tapes, GAM modified his program so that the data could be loaded in several sections, which was achieved satisfactorily.

The electron density was calculated at 1/120ths of the cell edge along *a* and *b*, and 1/60ths along *c*. The output from the computer was on punched paper tape, arranged with appropriate insertions of carriage returns and line feeds so that the teleprinter output was in a form suitable for immediate contouring to a scale of 0.75 inches equal to 1 Å. Each *x*, *y* section of the map was printed out in five strips, which had to be glued together to cover the whole area. The contours were drawn initially in pencil on these paper sheets and were then copied in drawing ink to thin Mylar sheets, which were supported for interpretation on Perspex sheets that were stacked and bolted together, with spacers of appropriate dimensions to maintain the scale in the *c* direction, in groups of five for ease of handling. The whole map was drawn on 60 sections perpendicular to the *z* axis, and the bolts holding the blocks of five sheets together were designed to fit into one another to keep successive blocks in register. The maps were viewed on large light boxes, specially constructed for the purpose, though they were not transparent enough for more than fifteen sheets, three blocks of five, to be studied in detail at one time. The contours were drawn on Mylar sheets to avoid waste of the more expensive Perspex that would have arisen from errors. A grid was drawn on a Mylar sheet, which could be superimposed on the contour stacks and used to read the atomic coordinates directly in ångstroms.

The electron density had been calculated using a scale factor such that it was convenient to draw contours at intervals of 0.25 e Å^{-3}; as no $F(000)$ term had been included in the calculation, electron-density values were relative to the mean value for the unit cell. Contours were drawn only for electron densities above the mean, the two lowest levels being drawn in orange ink, and the higher ones in black ink. This proved to be a satisfactory form of representation, as it very clearly revealed regions of continuous high electron density while also showing the positions of significant features of lower density. All the members of the lysozyme group participated in drawing the maps.

To illustrate the result that was obtained, sections $z = 35/60$ to 44/60 of the electron-density map are shown in Fig. 25.1.3.9.

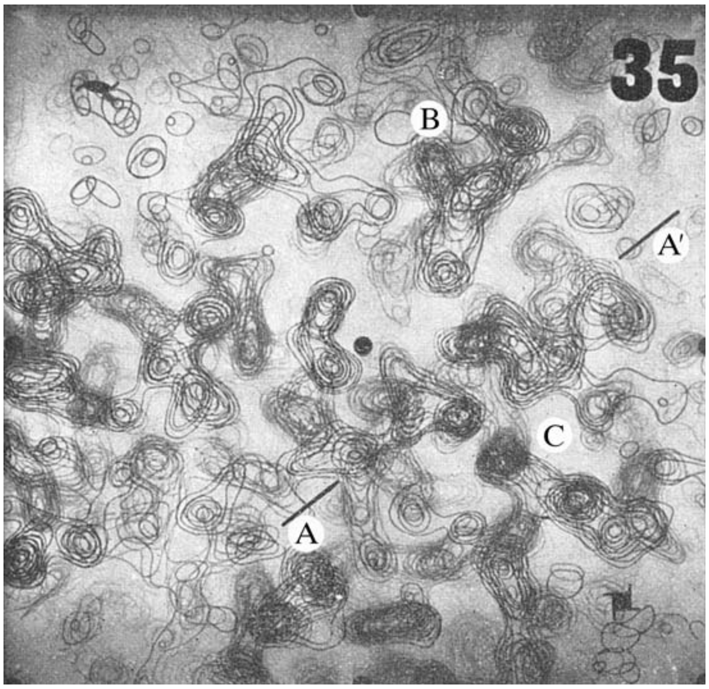


Figure 25.1.3.9
Photograph of sections $z = 35/60$ to 44/60 of the three-dimensional electron-density map of hen egg-white lysozyme at 2 Å resolution. AA′ shows the axis of a length of α-helix lying in the plane of the sections. B indicates an α-helix more nearly normal to the sections. C indicates the disulfide bridge between residues 30 and 115, the sulfur atoms of which lie one above the other. The side chain of a phenylalanine residue is located four residues along the helix from the disulfide, towards the lower sections. Reproduced with permission from *Nature* (Blake *et al.*, 1965). Copyright (1965) Macmillan Magazines Limited.

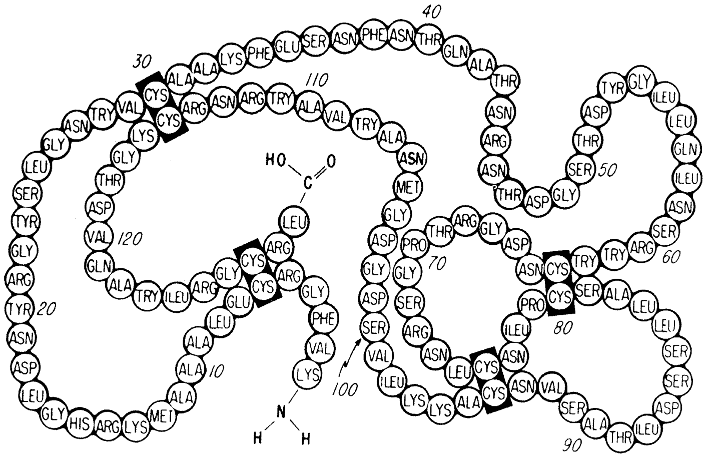


Figure 25.1.3.10
The amino-acid sequence of hen egg-white lysozyme (Canfield & Liu, 1965).

25.1.3.11. Map interpretation and model building

We were fortunate that by the time the map was ready for interpretation, two independent groups of protein chemists, led by Pierre Jollès in Paris and R. E. Canfield in New York, had studied the amino-acid sequence of hen egg-white lysozyme in detail and published their results. Two slightly different amino-acid sequences, complete with the arrangement of the four disulfide bonds, were published by Jollès *et al.* (1964) and by Canfield & Liu (1965), and this information was used intensively in the interpretation of the electron-density map of the protein. The sequence published by Canfield & Liu (1965) is shown in Fig. 25.1.3.10.

Many features were immediately identifiable in the map. These included the side chains of many of the amino-acid residues, especially the disulfide bridges and the aromatic side chains of tryptophan, tyrosine and phenylalanine. Nevertheless, no attempt was made to interpret the map in detail without recourse to the amino-acid sequence. Interpretation began in the part of the map shown in Fig. 25.1.3.9. The strongest feature in this part of the map corresponded to a disulfide bridge (C), in which the two sulfur atoms lie one above the other in the direction of the *c* axis. The first challenge was to identify this bridge. It is clearly connected to a helical region of the molecule (A), which ran in the direction from top right to bottom left of the diagram, with the main-chain carbonyl groups pointing in this direction. Consequently, this helix ran from its amino terminus on the right to its carboxyl terminus in the centre of the map. The map was quite clear enough to count the α-carbons from the cysteine residue that forms part of the disulfide bridge, and it was immediately apparent that the fourth residue from the cysteine towards the carboxyl terminus is an aromatic residue, probably phenylalanine. Inspection of the amino-acid sequence in Fig. 25.1.3.10 showed that only one pair of residues satisfied this condition, Cys30 and Phe34.

Given this start, interpretation of the map and the construction of a molecular model were relatively straightforward. The model was constructed in a metal frame, the top and bottom of which consisted of sheets of blockboard. The *a* and *b* axes were drawn parallel to the diagonals of these boards to cover the coordinate ranges, respectively, $x = -1/4$ to $+1/4$ and $y = 0$ to $+1/2$ to a scale of 2 cm to 1 Å. This was the scale of the brass models, constructed by Cambridge Repetition Engineers Ltd, which were used to build the model. The height of the frame covered the full extent of the *c* axis. The heavy-atom coordinates and the computer programs were both based on the wrong-handed space group $P4_12_12$. It was not until the anomalous scattering from the heavy atoms was incorporated that the correct space group $P4_32_12$ was

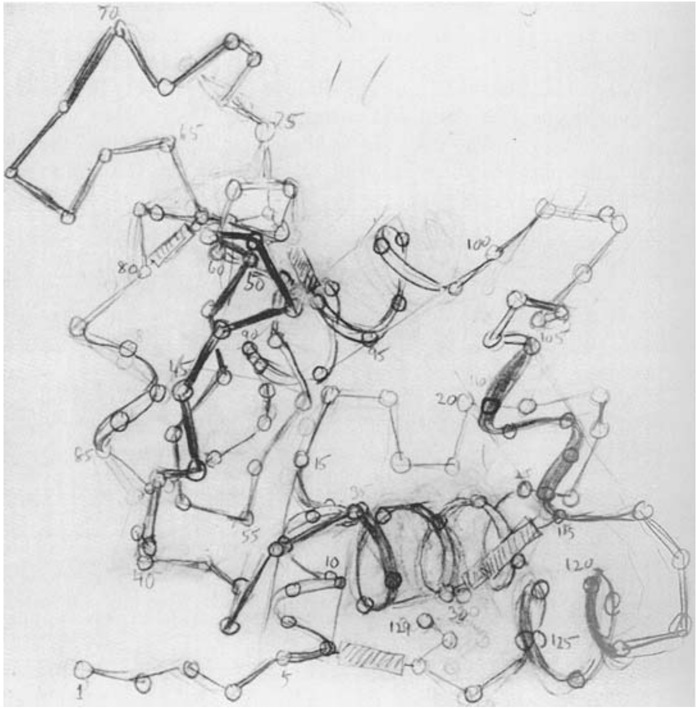


Figure 25.1.3.11
Schematic drawing of the main-chain conformation of lysozyme. The drawing was made from observations of the molecular model by Sir Lawrence Bragg and later prepared for publication by Mrs S. J. Cole.

Table 25.1.3.3
Discrepancies in amino-acid sequences (excluding Asp/Asn)

	Residue						
Reference	40	41	42	58	59	92	93
Canfield & Liu (1965)	Thr	Gln	Ala	Ile	Asn	Val	Asn
Jollès *et al.* (1964)	Gln	Ala	Thr	Asn	Ile	Asn	Val

assigned. The Fourier-map sheets were actually stacked the opposite way round and a left-handed system of axes was used for the model. In retrospect, this should have been put right at once, but the system was not easy to change.

Holes were drilled in the top and the base boards on the grid defined by the *a* and *b* axes, and these were used to support an array of brass rods parallel to the *c* axis to which the model components could be attached. The model building was carried out by two subgroups, CCFB and VRS in one and ACTN and DCP in the other, so that work could go on continuously throughout each day. The method employed was to examine the map density corresponding to the next amino-acid residue to be located and to mark the positions of the constituent atoms with small washers or nuts. The coordinates of these atoms were then read from the map (making use of the superimposed grid and estimating the *z* coordinates from the extent to which adjacent *z* sections contributed to the density). These coordinates were then located in the model by means of the coordinate grids drawn on the base and top boards and by the use of a plumb line marked with the *z* coordinates. At this stage it was usually possible to fix a model of the amino-acid residue in place in the model frame with remarkably little trouble, though, of course, fine adjustment was necessary as the model grew.

This bout of model building began towards the end of February 1965 and proceeded quite rapidly. The main difficulties arose from the fact that the two amino-acid sequences that were available did not agree in every respect. Eleven amino-acid residues were identified differently by Jollès (Jollès *et al.*, 1964) and by Canfield (Canfield & Liu, 1965). Four of the discrepancies involved Asn and Asp, which cannot be distinguished in the electron-density map with any degree of certainty. The remaining discrepancies are shown in Table 25.1.3.3.

Inspection of the residues 40, 41, 42, 92 and 93 showed quite clearly that the shapes in the electron-density map fitted the Canfield side chains. Our initial conclusion, however, was that the electron densities corresponding to residues 58 and 59 were more consistent with the sequence proposed by Jollès than that published by Canfield. Accordingly, in our first detailed description of the structure (Blake *et al.*, 1965), we accepted the identification of these residues by Jollès and his colleagues. As will be seen, this was a mistake.

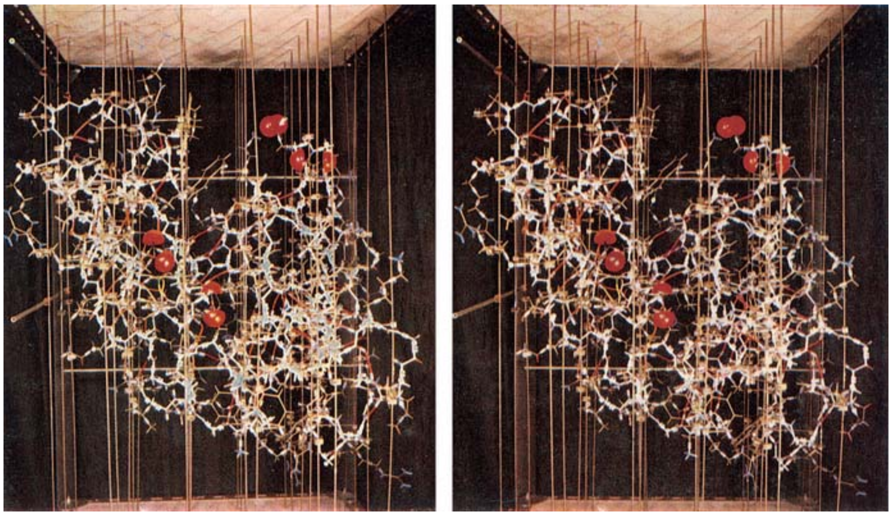


Figure 25.1.3.12
Stereo-photographs of a model of the lysozyme molecule to a scale of 2 cm to 1 Å. The main polypeptide chain is painted white, and nitrogen, oxygen and sulfur atoms are indicated by blue, red and green sleeving, respectively. Some hydrogen bonds are shown by red connections. Oxygen atoms of the acid side chains near the cleft, Glu 35, Asp 52, Asp 101 and Asp103, are shown by red hemispheres (Blake *et al.*, 1967).

The model was completed in time for it to form the centrepiece at Bragg's 75th birthday party at the Royal Institution on 31 March 1965, and a description was submitted to *Nature* at about the same time (Blake *et al.*, 1965). One of the difficulties at this stage of development of the subject was that computer graphics had not yet been developed to the stage at which illustrations of protein structures could be produced with any degree of facility. Consequently, we were most grateful that Sir Lawrence Bragg enthusiastically drew the model freehand. His drawing provided the main illustration for the published paper: the original sketch is reproduced in Fig. 25.1.3.11.

During the early summer of 1965, we rebuilt the model, taking care to ensure that the components fitted the electron density as well as could be judged by eye, and that the contacts between them were fully consistent with current understanding of van der Waals forces and hydrogen bonds. The model components were fixed firmly in position with small brass clamps, designed and constructed by Bruce

Morris in the Royal Institution workshop, so that at the end of the process we had an unrealistically rigid model that incorporated all the information available to us at the time. At this stage it became clear that the Canfield sequence (Fig. 25.1.3.10) was to be preferred in every respect.

A stereo-photograph of this model (Blake *et al.*, 1967) is shown in Fig. 25.1.3.12 in a view that shows the most striking feature of the molecule. The molecular face in the foreground of this view is crossed by a deep cleft, roughly parallel to the *c* axis of the crystal unit cell.

25.1.4. Structural studies on the biological function of lysozyme

25.1.4.1. Lysozyme substrates

By the summer of 1964, the results for the structure determination of lysozyme were sufficiently promising for us to consider diffraction studies on the biological function of lysozyme. At about the same time, the Royal Society invited Max Perutz to organise a Discussion Meeting on lysozyme at The Royal Institution in early 1966, to which the leading workers worldwide would be invited. This provided an additional spur to our efforts.

Simultaneously with his discovery of lysozyme in 1922, Fleming had found a gram-positive species of bacteria, *Micrococcus lysodeikticus*, that was particularly susceptible to lysis by lysozyme (Fleming, 1922). During the 1940s and 1950s, further work by a number of authors had shown that lysozyme exerted its biological effects through hydrolysis of the bacterial cell wall. These studies led to a definitive description of the polysaccharide component of the bacterial cell wall, as shown in Fig. 25.1.4.1.

The structure is composed of alternating sugars, linked by $\beta(1 \rightarrow 4)$ glycosidic links, of *N*-acetylglucosamine (GlcNAc, abbreviated by us at the time as NAG) and *N*-acetylmuramic acid (MurNAc, previously abbreviated as NAM). The 3-hydroxyls of the MurNAc residues are attached to short peptides, and the peptides themselves are cross-linked to provide an extensive and rigid two-dimensional proteoglycan network. Lysozyme cleaves the $\beta(1 \rightarrow 4)$ glycosidic bonds between MurNAc and GlcNAc residues, thus leading to the dissolution of the bacterial coat and lysis of the bacterium.

Kinetic studies on the activity of lysozyme were hampered by the lack of a suitable small-molecular-weight substrate. The turbidometric assay in use at the time, and still in use today, followed the change in optical density of a suspension of *Micrococcus lysodeikticus* cells as the cells were lysed by lysozyme. The assay could work reliably but it was sensitive to physical parameters such as ionic strength, the method by which the cells were suspended and product inhibition. In 1962, Wenzel *et al.*, in an effort to obtain a small-molecular-weight substrate, had reported that lysozyme promoted the cleavage of the trimer of GlcNAc, tri-*N*-acetylchitotriose, releasing dimer and monomer sugars, and that the monomer, GlcNAc, was an inhibitor of lysozyme (Wenzel *et al.*, 1962). Compounds such as glucose or cellobiose that lacked the *N*-acetyl group did not inhibit.

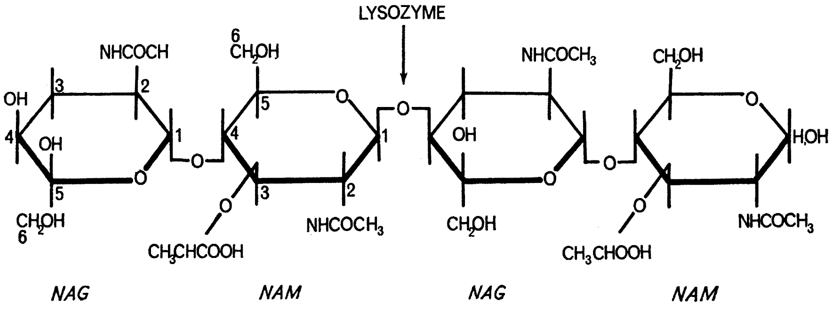


Figure 25.1.4.1
The cell-wall tetrasaccharide with the $\beta(1 \rightarrow 4)$ glycosidic bond that is hydrolysed by lysozyme indicated (Blake *et al.*, 1967).

John Rupley extended this work. Rupley, a chemist from the University of Arizona, Tucson, played a crucial role in developing the mechanisms for lysozyme catalysis. In results that were summarized in 1964 and 1967 (Rupley, 1964, 1967) and made available to workers at the Royal Institution in 1965, he showed that the rate of hydrolysis of GlcNAc homopolymers increased by 30 000 as the chain length increased from trimer to hexamer, and that there was no further increase in rate for substrates of greater length than the hexamer. Furthermore, the cleavage patterns for the smaller polymers were complex, but the optimal substrate, the hexamer, was cleaved to a tetrasaccharide and a disaccharide.

Thus by the time we began structural studies, the nature of the substrate and the specificity of lysozyme were established. GlcNAc was the only compound that was available commercially. This seemed a good starting point. It was anticipated that as a competitive inhibitor it would bind to and allow the identification of the catalytic site.

25.1.4.2. The crystal structure of GlcNAc

During the first two years of her graduate work (1962–1964), LNJ determined the crystal structure of GlcNAc. β-D-*N*-Acetylglucosamine was crystallized from a methanol–water mixture, and diffraction data were recorded on film with a modified Weissenberg camera. Intensities were estimated visually from comparison with a scale of fixed-time exposures of the attenuated main beam. The structure was solved manually using a sharpened Patterson function and application of the minimum function of Buerger (1959). The structure revealed a standard glucopyranose ring in the chair conformation with the plane of the *N*-acetyl group normal to the ring (Johnson & Phillips, 1964; Johnson, 1966). The structure of glucosamine hydrochloride had been solved by Cox & Jeffrey in 1939 (Cox & Jeffrey, 1939), a remarkable early achievement.

By 1963, about ten glucopyranose structures were available, including that of cellobiose (Jacobson *et al.*, 1961). An analysis by Ramachandran *et al.* (1963) showed that the glucopyranose ring is remarkably uniform in its conformation and may be regarded as a rigid structure.

The final difference-Fourier synthesis of β-GlcNAc revealed an additional peak adjacent to the C1 carbon atom in the position of a hydroxyl group in the α configuration. Peak heights suggested that there could be a mixture in the crystal of approximately 80% β-*N*-acetylglucosamine and 20% α-*N*-acetylglucosamine. Refinement indicated that a mixture of the two anomers could be accommodated in the crystal lattice.

Tests on the optical rotation of the crystalline sample after dissolution, compared with the starting material, also added support to the notion that the crystal contained a mixture of α and β anomers, probably as a result of mutarotation during crystallization. Although a later and more precise structure determination of GlcNAc indicated a lower proportion of the α anomer (Mo & Jensen, 1975), mixtures of α and β sugars in crystals have been observed for other compounds (Jeffrey, 1990). The consideration of the α and β anomers of GlcNAc meant that we were keenly aware of the importance of configuration at the C1 atom. This turned out to be essential when we were interpreting the results with lysozyme.

25.1.4.3. Low-resolution binding studies of lysozyme with GlcNAc and other sugars

In the first binding studies of lysozyme with GlcNAc, crystals were soaked overnight in solutions of 0.15 *M* GlcNAc in the standard crystallization medium, and a 15° precession photograph was recorded by LNJ. The changes in intensities of reflections compared to a native lysozyme photograph were extremely small, much smaller than those observed with the heavy-atom derivatives, but were sufficiently promising to encourage us to collect three-dimensional data to 6 Å resolution. Data were collected on the linear diffractometer using the three counters in the flat-cone setting over a period of about 20 h, as described for the heavy-atom derivatives. The data were processed using the programs of ACTN and VRS on the Elliott 803B computer. The 6 Å difference-Fourier map, using the phases from the improved set of heavy-atom derivatives, was obtained at the end of October 1964. It showed a single rather elongated peak which, when superimposed onto the 6 Å model, was located in the cleft in the enzyme surface between the two domains (Johnson & Phillips, 1965).

The power of the difference-Fourier technique in protein crystallography was immediately demonstrated by this first 6 Å electron-density difference map for the lysozyme–inhibitor complex, as had also been demonstrated earlier in the work (Stryer *et al.*, 1964) on the binding of azide to sperm-whale myoglobin. Once a protein structure had been solved, it was apparent that ligand-binding sites could be established with ease. Following the GlcNAc result, 6 Å binding studies were repeated with a number of other compounds (Blake *et al.*, 1967). Kinetic studies using the turbidometric assay were carried out with each of the compounds in order to establish the mode of inhibition of lysozyme activity. We were fortunate in being able to use the skills of JWHO, a carbohydrate chemist who had been at the Royal Institution for many years. JWHO had synthesized 6-iodo-6-deoxy-*N*-acetyl β-methylglucosaminide, a compound which was found to inhibit more powerfully than GlcNAc itself. The low-resolution binding study showed a stronger and more compact peak at the catalytic site than that observed with GlcNAc, but it was not possible to resolve the iodine and hence identify the six positions of the sugar.

By January 1965, we had been given a sample of the disaccharide di-*N*-acetylchitobiose, $(\mathrm{GlcNAc})_2$, sent by John Rupley. Efficiency of data collection and map production had increased. Data collection was started on 19 January and the map was obtained by 28 January. Other results followed with *N*-acetylmuramic acid (a gift from R. W. Jeanloz) and the disaccharide *N*-acetylglucosamine β-(1,4)-*N*-acetylmuramic acid (GlcNAc-MurNAc – a gift from N. Sharon) (Fig. 25.1.4.2).

As part of these studies, penicillin V and *p*-iodophenoxymethyl penicillin potassium salt were also investigated. They were observed to inhibit lysozyme. Crystallographic studies showed that the penicillins did indeed bind to lysozyme in the catalytic cleft, but at a site remote from the GlcNAc binding site (Johnson, 1967). As established shortly after this result was obtained in 1965, penicillin exerts its potent antibiotic activity by inhibition of the enzymes responsible for the biosynthesis of the peptide cross-linking component of the bacterial cell wall (Tipper & Strominger, 1965; Wise & Park, 1965). An original suggestion that penicillin might resemble MurNAc turned out to be incorrect (Collins & Richmond, 1962). The interactions of penicillin with lysozyme are probably fortuitous, but were not fully investigated.

By the end of May 1965, we were ready to move on to 2 Å data collection, a formidable task that required fourteen crystals and more than two weeks continuous data collection on the multiple-counter linear diffractometer. Data for the lysozyme–GlcNAc complex were completed first, and the map was available around October 1965. The electron density was puzzling, and an interpretation was not possible until the high-resolution results with the trisaccharide were available.

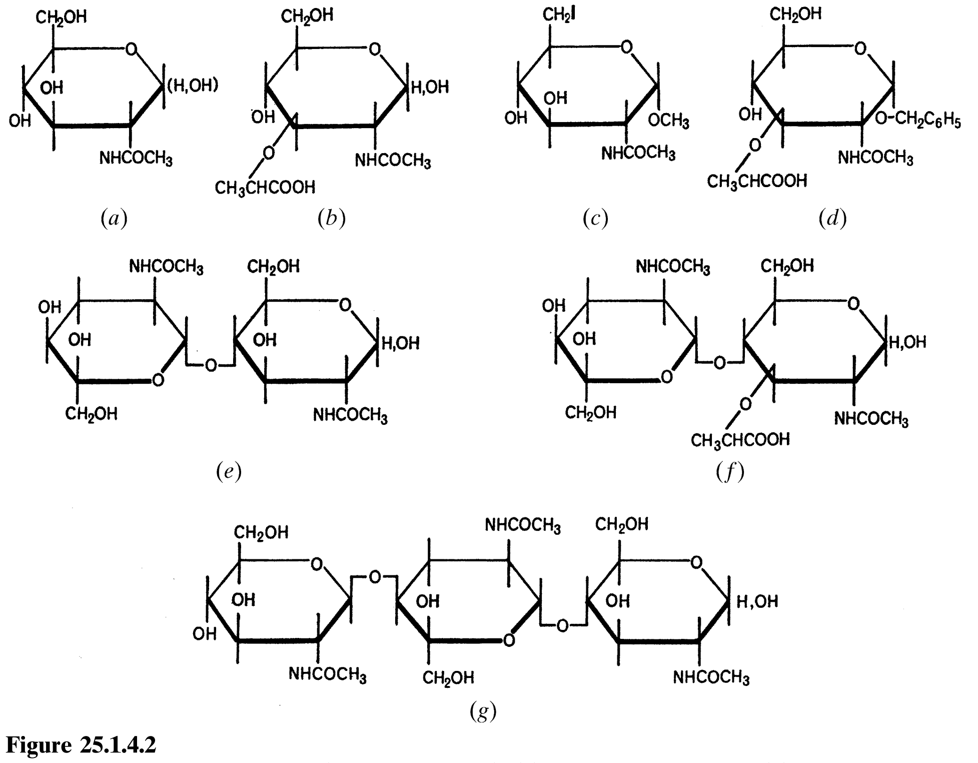


Figure 25.1.4.2

Inhibitor molecules of lysozyme (Blake *et al.*, 1967). (*a*) *N*-acetylglucosamine; (*b*) *N*-acetylmuramic acid; (*c*) 6-iodo-α-methyl-*N*-acetylglucosaminide; (*d*) α-benzyl-*N*-acetylmuramic acid; (*e*) di-*N*-acetylchitobiose; (*f*) *N*-acetylglucosaminyl-*N*-acetylmuramic acid; (*g*) tri-*N*-acetylchitotriose.

25.1.4.4. Binding studies of lysozyme with tri-N-acetyl-chitotriose, $(\mathrm{GlcNAc})_3$, at 2 Å resolution

John Rupley visited the Royal Institution in July 1965. He brought with him a sample of tri-*N*-acetylchitotriose (Fig. 25.1.4.2), previously abbreviated as tri-NAG and now as $(\mathrm{GlcNAc})_3$. The lysozyme–$(\mathrm{GlcNAc})_3$ complex turned out to be the crucial structure for understanding activity. When lysozyme crystals were soaked in 0.05 *M* $(\mathrm{GlcNAc})_3$ solution, they turned opaque and became difficult to mount for X-ray studies. The problem was overcome by soaking the crystals already mounted in a capillary tube. By 17 July 1965, a 6 Å map of the lysozyme–$(\mathrm{GlcNAc})_3$ complex was obtained and showed a peak occupying the whole top part of the cleft on the enzyme surface.

Rupley had also succeeded in co-crystallization of lysozyme with $(\mathrm{GlcNAc})_3$. Data collected from these crystals gave an identical difference-Fourier synthesis to that obtained with crystals soaked in $(\mathrm{GlcNAc})_3$. Data from another co-crystallized crystal, which were collected ten days later, also gave an identical result, indicating that there had been no hydrolysis of the trisaccharide in the crystal during this time. The way was open for high-resolution data collection with $(\mathrm{GlcNAc})_3$. Rupley set up co-crystallizations of lysozyme with $(\mathrm{GlcNAc})_3$ (with concentrations 1:1.1 molar ratio under standard conditions) at the Royal Institution during the hot August of 1965. The crystals grew well and provided us with a plentiful supply of crystals for the 2 Å data collection, which was carried out by LNJ following the same procedures as for the native crystals described above.

The 2 Å difference-electron-density map arrived on Thursday 6 January 1966, but there was a mistake. After correction and the laborious task of contouring by hand and transferring the contours to transparent sheets had been completed, serious model building by DCP, LNJ, ACTN and other members of the team began on 19 January 1966.

The interpretation of the difference map for $(\mathrm{GlcNAc})_3$ was straightforward. Three sugars could readily be identified in the density. The first occupied a site at the centre of the cleft at a position similar to that occupied by GlcNAc by itself and labelled site C. The acetamido group was visible and fitted neatly into a pocket where there were complementary hydrogen bonds between the NH and CO of the acetamido group and the main-chain carbonyl oxygen of residue 107 and the amido group of residue 59. Lysozyme has three tryptophan residues at the catalytic site. One of these, Trp108, was at the bottom of the acetamido pocket and made van der Waals contacts to the methyl group of the acetamido. The precise and extensive contacts to the acetamido group explained immediately the specificity of lysozyme for *N*-acetyl sugars. The free reducing group of the sugar in site C pointed down (towards lower z). The second sugar could be fitted to the density above the first (in the direction of increasing z), linked to the O4 of the sugar in site C. It was clear that a second tryptophan, Trp62, stacked against the glucopyranose ring for this sugar in site B, and that there was a shift of this tryptophan towards the sugar. This and other conformational changes could explain the tendency of the lysozyme crystals to become disordered when soaked in $(\mathrm{GlcNAc})_3$ solutions.

In cellobiose, whose crystal structure had been determined in 1961 by Lipscomb and colleagues (Jacobson *et al.*, 1961), the conformation about the $\beta(1\rightarrow4)$ glycosidic link rotates the second sugar about 180° with respect to the first, and there is an intramolecular hydrogen bond between the O3 hydroxyl of one sugar and the O5 of the adjacent sugar. This relative orientation of the two adjacent sugars was found to fit the density for sugars C and B and also to agree with models proposed by Carlstrom (1962) for chitin and the analysis of Ramachandran *et al.* (1963). The lobes representing the acetamido and C6 alcohol groups for site B provided further confirmation for this orientation. The third sugar in site A appeared less well located, but nevertheless could be placed in density and was observed to make satisfactory contacts with the protein.

The interpretation of the difference-Fourier map was made by one worker without reference to the protein structure, while a second worker fitted the model into the protein structure. It was most exciting and satisfying to observe the contacts when the skeletal models of the sugars were fitted into the skeletal model of the protein. Shifts in atomic coordinates of selected protein groups were estimated from the gradient of electron density in a difference map divided by the curvature of the electron density at that point, as described by Lipson & Cochran (1968). Using this formula, the shift in Trp62 was calculated to be 0.75 Å, a value that turned out to be accurate when the lysozyme–$(\mathrm{GlcNAc})_3$ complex was refined by least-squares methods many years later (Cheetham *et al.*, 1992). It was also noted that there were two acidic residues on either side of the catalytic cleft, Glu35 and Asp52, which were some distance from the $(\mathrm{GlcNAc})_3$ position. There was a further acidic residue, Asp101, near the top of the cleft.

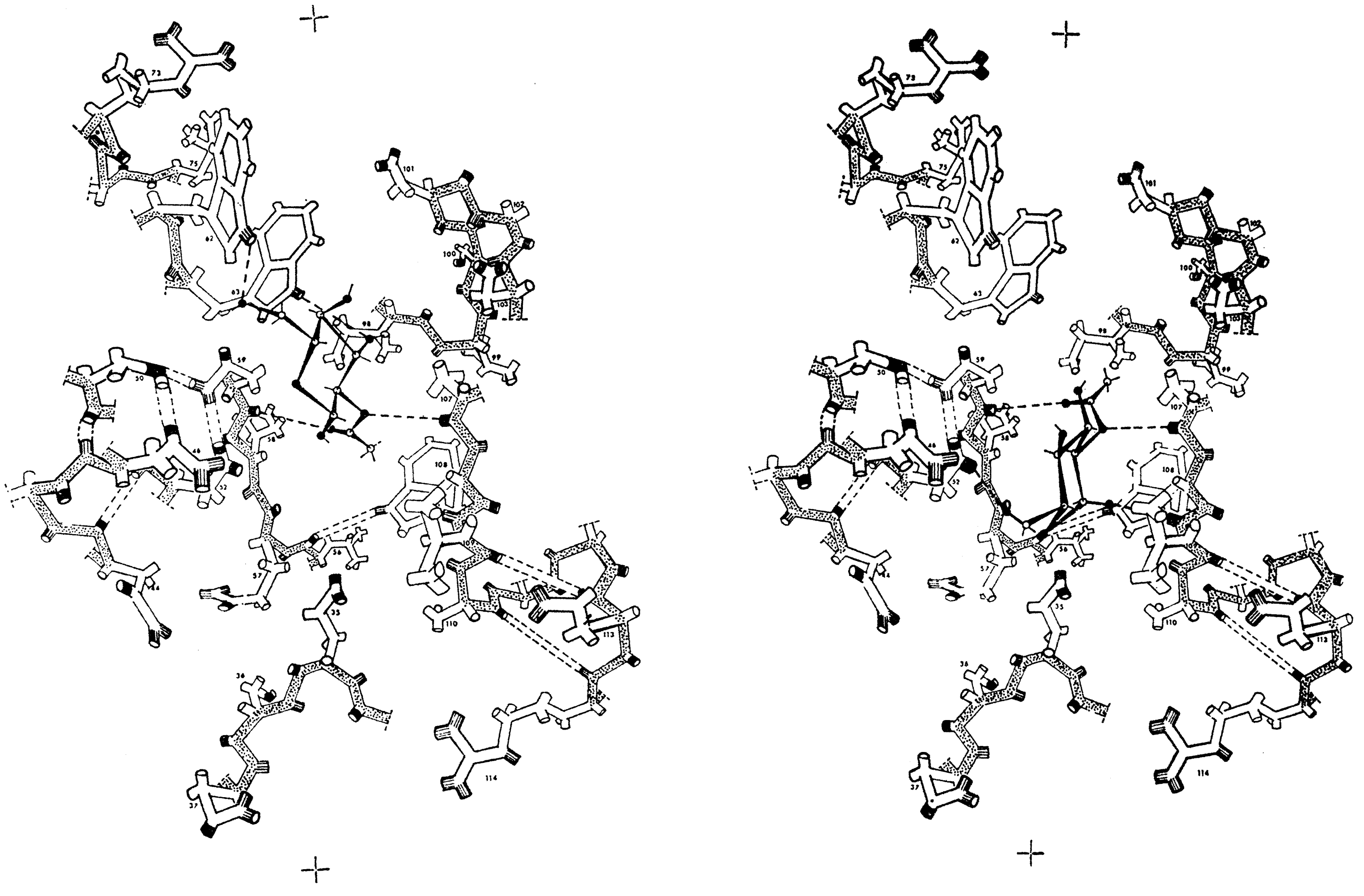
The structure of the $(\mathrm{GlcNAc})_3$–lysozyme complex allowed an interpretation of the GlcNAc result. It was apparent that GlcNAc bound to site C in one or other of two distinct but closely related ways, depending upon whether it is in the α or β enantiomeric form (Fig. 25.1.4.3).

The GlcNAc molecule had evidently undergone mutarotation in solution to produce an equilibrium mixture of α and β forms, both of which bound to lysozyme, although of course the specificity of the enzyme for longer oligosaccharides is for β-linked sugars. Both binding modes exploited the specificity of the *N*-acetyl group and its interactions with the enzyme.

One mode was characteristic of β-GlcNAc, as observed for the terminal sugar of $(\mathrm{GlcNAc})_3$. The other mode was characteristic of α-GlcNAc and involved a rotation about the hydrogen bonds of the acetamido group to the protein. This finding could explain the result for 6-iodo-6-deoxy-*N*-acetyl α-methylglucosaminide, which could not fit into the α-GlcNAc site because of the additional methyl group but could only be accommodated in the 'β-site'. The dual mode of binding for the terminal residue in the α or β configuration could also explain the bifurcated peak observed for the disaccharides GlcNAc-MurNAc and $(\mathrm{GlcNAc})_2$. When the terminal sugar is in the α configuration, the second sugar is placed out of the main run of the cleft, while in the β configuration the second residue is as in site B of the trisaccharide complex.

25.1.4.5. Proposals for the catalytic mechanism of lysozyme

A Royal Society Discussion Meeting had been arranged for 3 February 1966 under the organization of Max Perutz. CAV, from University College, London, who had worked on mechanisms for the non-enzymic hydrolysis of glycosides in solution, had been invited to contribute to this meeting. In mid-January (the exact date is not recorded) he visited the Royal Institution to discuss the possible reaction mechanism of lysozyme in the light of the crystallographic evidence. At that stage, the crystallography had shown, from low-resolution studies, the inhibitor binding sites within a surface cleft lined with several acidic residues that included Glu35 and Asp52. CAV's work had shown that the hydrolysis of methyl-α-D-glucopyranoside proceeded *via* a ring-closed carbonium-ion intermediate (*i.e.* a carbocation) and that the rate-determining step was the heterolysis of the C1 carbon–oxygen bond (reviewed in Vernon, 1967). Rupley's studies with ^{18}O water had shown that hydrolysis of oligosaccharide substrates by lysozyme also proceeded through cleavage of the C1—O bond (Rupley, 1967). Whereas in free solution the carbocation intermediate is presumed to be stabilized by interaction with solvent, in the catalytic site of lysozyme it was suggested that the carbonium-ion intermediate might be stabilized by the nearby ionized carboxylate of one of the acidic residues in the catalytic site.



(*a*) (*b*)

Figure 25.1.4.3
Contemporary drawings of the binding to lysozyme of: (*a*) β-*N*-acetylglucosamine and (*b*) α-*N*-acetylglucosamine (Blake *et al.*, 1967).

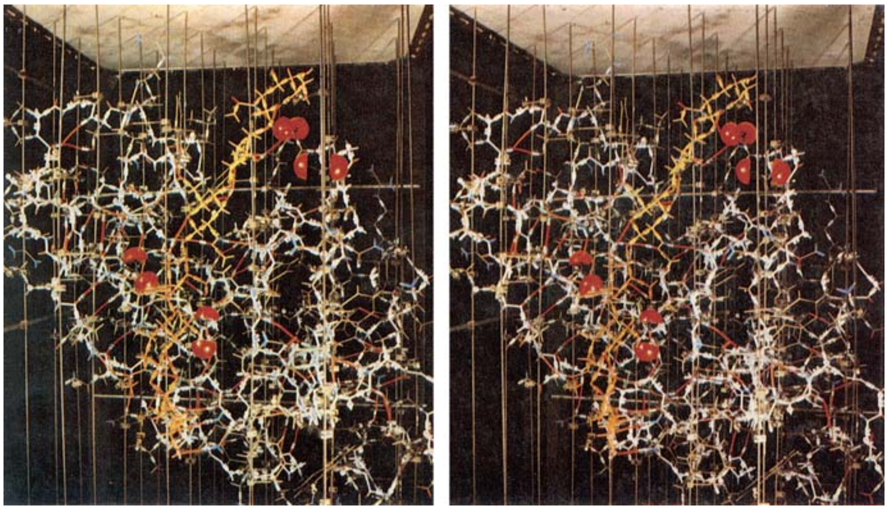


Figure 25.1.4.4
Stereo-photographs of a model of the lysozyme molecule showing how a hexasaccharide substrate may bind to the enzyme. The yellow trisaccharide model shown at the top of the cleft is in the position occupied by tri-*N*-acetylchitotriose in the crystals of its complex with lysozyme. The orange trisaccharide in the lower part of the cleft has been added by model building alone. The fourth sugar residue from the top, the uppermost orange one, is distorted from the chair conformation (Blake *et al.*, 1967). The glycosidic linkage that is broken is the one between the fourth and fifth sugar residues from the top. It lies between the side chains of Glu35 (to the right) and Asp52 (to the left), the oxygen atoms of which are marked by red hemispheres. Asp101 and Asp103, near the top of the cleft, also have their oxygen atoms marked.

The interpretations of the $(\mathrm{GlcNAc})_3$ and the GlcNAc complexes were finished by Monday 31 January 1966. The date is significant. The Royal Society Discussion Meeting had been arranged for the following Thursday. The results with the non-hydrolysable complex with $(\mathrm{GlcNAc})_3$ had identified the catalytic site and had provided immediate explanations for the specificity requirement for the enzyme. However, information on how a true substrate might bind was missing. This problem was solved in one day. DCP, noting that Rupley's work had shown the optimal substrate to be a hexasaccharide, constructed a hexasaccharide substrate in sites A–F, using the position of the experimentally determined trisaccharide in sites A–C and model building those sugars in sites D–F. Noting also the specificity of lysozyme for bacterial-cell-wall substrates, where the polysaccharide is composed of alternating $\beta(1\rightarrow4)$ linked GlcNAc and MurNAc residues, and the bond cleaved is that between MurNAc and GlcNAc residues, but not that between GlcNAc and MurNAc residues, he also examined sites likely to be specific for MurNAc. It was apparent that, for steric reasons, sites A, C and E could not accommodate MurNAc residues, but that sites B, D and F could do so. Thus the site of cleavage must be between sites D and E (or sites A and B, but this could be discarded because the trisaccharide complex was stable).

Close to the bond between sites D and E were two acid residues (Fig. 25.1.4.4). As noted previously by members of the team, one of these, Glu35, was in a non-polar environment and shielded by Trp108, Val109 and Ala110. In this environment, the carboxylic acid was likely to have a raised p*K* and to be protonated at pH 5, the optimal pH for activity against homopolymers of GlcNAc. The other acid group, Asp52, was also buried, but in an environment that was largely polar and at the centre of a hydrogen-bonding network. Hence it might be predicted that its p*K* was likely to be similar to that of a carboxyl group in an aqueous environment, and the group would therefore be mostly ionized at pH 5. There was one other important factor. The sugar in site D could not be accommodated in the usual chair conformation of a glucopyranose ring. In order to avoid overcrowding between the C6O6 atoms of the alcohol group and the protein atoms at site D, it was necessary to distort the sugar ring from the conventional chair conformation to a sofa conformation. This relieved the overcrowding by bringing the C5—C6 bond axial and retained the links from C4—O4 and C1—O4′ equatorial, so that the adjacent sites of C and E were not perturbed by the distortion of the sugar in site D (Fig. 25.1.4.4). CAV visited the Royal Institution on a second occasion between the completion of the model and the Royal Society meeting, probably on 1 February. When shown the hexasaccharide substrate complex with two acid residues on either side of the susceptible bond and the distortion of the sugar in site D, CAV explained that this was exactly what he would expect from his own work for a carbonium-ion mechanism with transition-state distortion of the substrate.

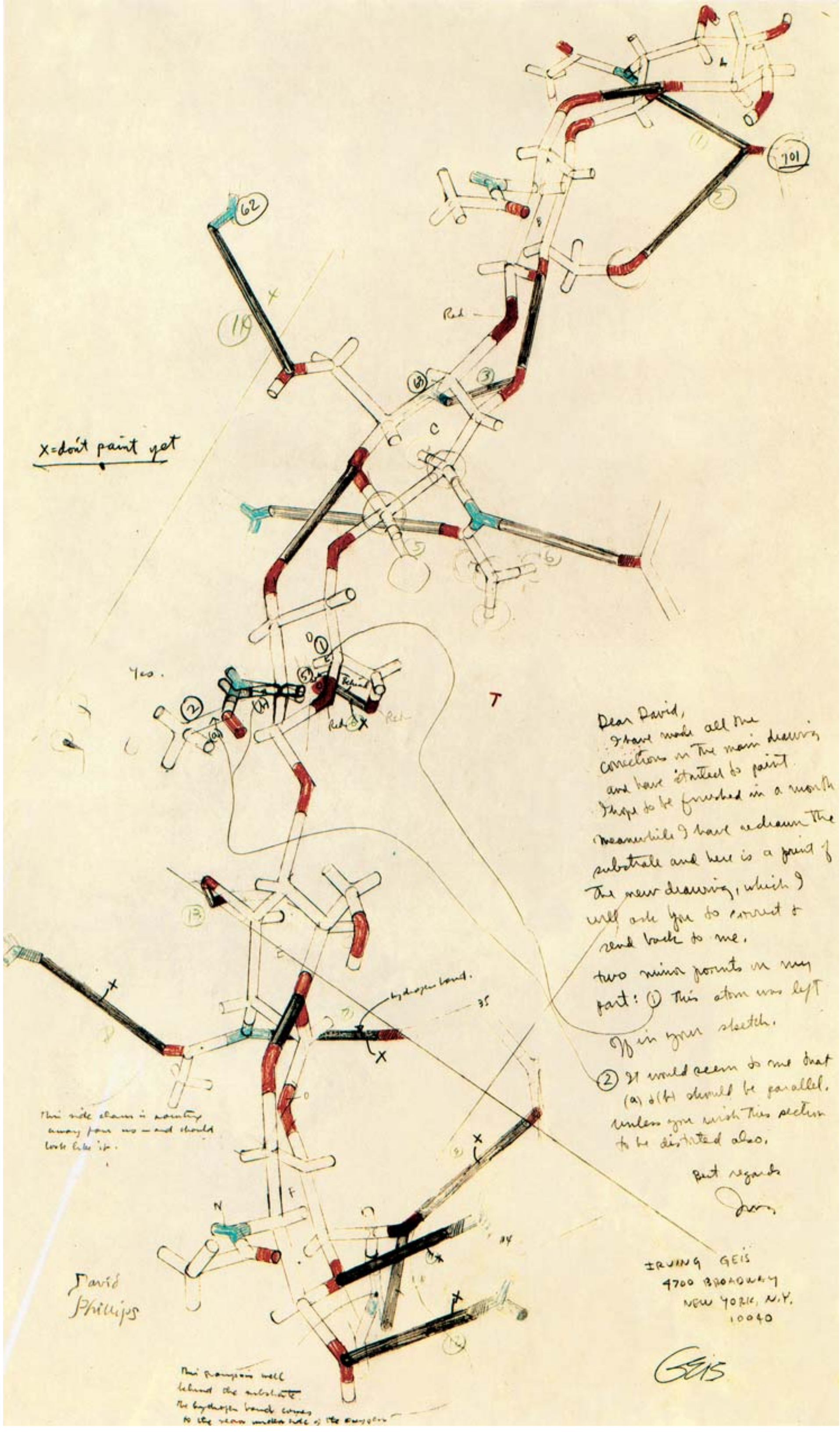

101
62
x
x=don't paint yet
C
Red
T
Dear David,
I have made all the corrections in the main drawing and have started to paint. I hope to be finished in a month.
Meanwhile I have redrawn the substrate and here is a print of the new drawing, which I will ask you to correct & send back to me.
two minor points in my part: ① This atom was left off in your sketch.
② It would seem to me that (a) & (b) should be parallel, unless you wish this section to be distorted also.
Best regards
Irv
IRVING GEIS
4700 BROADWAY
NEW YORK, N.Y.
10040
GEIS
hydrogen bond.
35
x
This side chain is pointing away from us — and should look like it.
N
F
O
34
David Phillips
This oxygen well behind the substrate
The hydrogen bond comes to the rear underside of the oxygen

Figure 25.1.4.5
Draft sketch of the lysozyme–hexasaccharide substrate complex prepared by Irving Geis and annotated by David Phillips for an article published in *Scientific American* in 1966 (Phillips, 1966).

The mechanism was presented for the first time at the Royal Society meeting on 3 February 1966. It was noted that the nearest oxygen of the protonated Glu35 was about 3 Å from the glycosidic oxygen between residues D and E (the O1 atom of residue D and the equivalent O4 atom of residue E). On the other side of the cleft, the nearest oxygen of the ionized Asp52 was about 3 Å from the C1 atom of residue D and about the same distance from the O5 of that residue. The distances were, of course, subject to some error from experimental model building. It was suggested (Blake *et al.*, 1967; Phillips, 1967) that:

(1) Asp52 carries a negative charge that promotes the formation of a carbonium ion at C1 of residue D and stabilizes it when formed;
(2) distortion of residue D from the chair conformation into the sofa conformation would contribute to stabilization of a carbonium ion at C1 by favouring a conformation in which the charge at C1 could be shared with the ring oxygen atom (Lemieux & Huber, 1955), and hence

contribute to the consequent weakening of the C1—O1 bond; and

(3) Glu35 could act as a proton donor, facilitating the formation of a hydroxyl group with the bridge oxygen atom (O1), and release of residues E and F.

Fig. 25.1.4.5 shows a sketch of the proposed lysozyme–hexasaccharide substrate complex prepared by Irving Geis for a *Scientific American* article published later that year (Phillips, 1966).

As Perutz commented in his closing remarks at the Royal Society meeting: 'For the first time we have been able to interpret the catalytic activity of an enzyme in stereochemical terms' (Perutz, 1967). The scheme for catalysis has been subjected to numerous experimental tests and has been investigated further in crystallographic experiments. The mechanism provided a satisfactory explanation for the body of evidence already in existence at the time of the proposals and has been substantiated by the results of new experiments designed to test it (see, for example, Imoto *et al.*, 1972; Ford *et al.*, 1974; Strynadka & James, 1991; Hadfield *et al.*, 1994; Jolles, 1996).

This work would not have been possible without the continuous support and encouragement of Sir Lawrence Bragg and the Medical Research Council, UK, for which we were and are most grateful. We are also grateful to our colleagues Robert Canfield, Pierre Jolles, Gordon Lowe, John A. Rupley, Nathan Sharon and others who shared their results on the catalytic mechanism with us before publication. The successful outcome also depended greatly upon the skilled work of the craftsmen in the Royal Institution's mechanical and electronic workshops, Messrs T. H. Faulkner, S. B. Morris, J. E. T. Thirkell and A. R. Knott. Finally, we acknowledge gratefully the unstinting contributions made by our team of research assistants, Mrs W. J. Browne, Mrs A. Hartley, Mrs K. Sarma, Miss D. Glass, Mrs R. Arthanari, Mrs S. J. Cole, Mrs J. A. Conisbee and Miss M. Hibbs. Clearly, thirty years ago protein-structure analysis was still very labour-intensive, despite the advances in computing and diffractometry described above. This was largely because we were still designing and building our own apparatus and drawing maps and building models by hand. This situation has been transformed chiefly by the availability of more powerful computers and X-ray detectors, by the developments of computer graphics, and by the exploitation of synchrotron radiation. The review has used the original figures prepared during the 1960s with a few exceptions. We acknowledge with special thanks the skills of Stephen Lee in scanning and reproducing these figures.

References

Abraham, E. P. & Robinson, R. (1937). *The crystallization of lysozyme. Nature (London)*, **140**, 24.

Alderton, G., Ward, W. H. & Fevold, J. (1945). *Crystallization of lysozyme chloride. J. Biol. Chem.* **157**, 43–58.

Arndt, U. W., North, A. C. T. & Phillips, D. C. (1964). *Adaptation of a linear diffractometer to measure three X-ray reflections quasi-simultaneously. J. Sci. Instrum.* **41**, 421–425.

Arndt, U. W. & Phillips, D. C. (1961). *The linear diffractometer. Acta Cryst.* **14**, 807–818.

Arndt, U. W. & Riley, D. P. (1952). *Side-window proportional counters. Proc. Phys. Soc. London A*, **65**, 74–84.

Bijvoet, J. M. (1954). *Structure of optically active compounds in the solid state. Nature (London)*, **173**, 888–891.

Blake, C. C. F. (1968). *The preparation of isomorphous derivatives. Adv. Protein Chem.* **23**, 59–120.

Blake, C. C. F., Fenn, R. H., North, A. C. T., Phillips, D. C. & Poljak, R. J. (1962). *Structure of lysozyme. A Fourier map of the electron density at 6 Å resolution obtained by X-ray diffraction. Nature (London)*, **196**, 1173–1178.

Blake, C. C. F., Johnson, L. N., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Crystallographic studies of the activity of hen egg-white lysozyme. Proc. R. Soc. London Ser. B*, **167**, 378–388.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **206**, 757–761.

Blake, C. C. F. & Phillips, D. C. (1962). *Effects of X-irradiation on single crystals of myoglobin*. In *Biological Effects of Ionizing Radiation at the Molecular Level*. Brno, Czechoslovakia: International Atomic Energy Agency, Vienna.

Blow, D. M. (1958). *The structure of haemoblobin. VII. Determination of phase angles in the non-centrosymmetric [100] zone. Proc. R. Soc. London Ser. A*, **247**, 302–336.

Blow, D. M. & Crick, F. H. C. (1959). *The treatment of errors in the isomorphous replacement method. Acta Cryst.* **12**, 794–802.

Blow, D. M. & Rossmann, M. G. (1961). *The single isomorphous replacement method. Acta Cryst.* **14**, 1195–1202.

Bluhm, M. M., Bodo, G., Dintzis, H. M. & Kendrew, J. C. (1958). *The crystal structure of myoglobin IV. A Fourier projection of sperm-whale myoglobin by the method of isomorphous replacement. Proc. R. Soc. London Ser. A*, **246**, 369–389.

Buerger, M. J. (1959). *Vector Space*. New York: Wiley.

Canfield, R. E. & Liu, A. K. (1965). *The disulfide bonds of egg white lysozyme (muramidase). J. Biol. Chem.* **240**, 1997–2002.

Carlstrom, D. (1962). *The polysaccharide chain of chitin. Biochim. Biophys. Acta*, **59**, 361–364.

Cheetham, J. C., Artymiuk, P. J. & Phillips, D. C. (1992). *Refinement of an enzyme complex with inhibitor bound at partial occupancy. J. Mol. Biol.* **224**, 613–628.

Collins, J. F. & Richmond, M. (1962). *A structural similarity between N-acetyl muramic acid and penicillin as a basis for antibiotic action. Nature (London)*, **195**, 142–143.

Corey, R. B., Donohue, J., Trueblood, K. N. & Palmer, K. J. (1952). *An X-ray investigation of air-dried lysozyme chloride crystals: the three-dimensional Patterson function. Acta Cryst.* **5**, 701–710.

Cox, E. G. & Jeffrey, G. A. (1939). *Crystal structure of glucosamine hydrobromide. Nature (London)*, **143**, 894–895.

Crick, F. H. C. & Magdoff, B. S. (1956). *The theory of the method of isomorphous replacement for protein crystals. 1. Acta Cryst.* **9**, 901–908.

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1961). *The structure of haemoglobin VIII. A three-dimensional Fourier synthesis at 5.5 Å resolution: determination of the phase angles. Proc. R. Soc. London Ser. A*, **265**, 15–38.

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G. & North, A. C. T. (1962). *The structure of haemoglobin IX. A three-dimensional Fourier synthesis at 5.5 Å resolution: description of the structure. Proc. R. Soc. London Ser. A*, **265**, 161–187.

Dauben, C. H. & Templeton, D. H. (1955). *A table of dispersion corrections for X-ray scattering of atoms. Acta Cryst.* **8**, 841–842.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *The crystal structure of myoglobin: phase determination to a resolution of 2 Å by the method of isomorphous replacement. Acta Cryst.* **14**, 1188–1195.

Dickerson, R. E., Reddy, J. M., Pinkerton, M. & Steinrauf, L. K. (1962). *A 6 Å model of triclinic lysozyme. Nature (London)*, **196**, 1178.

Fenn, R. H. (1964). *An X-ray Crystallographic Study of Some Mercury Compounds and Their Use in Protein Structure Analysis.* Davy Faraday Research Laboratory, The Royal Institution, London 182.

Fenn, R. H., Phillips, D. C. & Oldham, J. W. H. (1963). *Crystal structure of* $(CH_3)_3S.HgI_3$ *and the configuration of the* HgI_3^- *ion. Nature (London)*, **198**, 381–382.

Fleming, A. (1922). *On a remarkable bacteriolytic element found in tissues and secretions. Proc. R. Soc. London Ser. B*, **93**, 306–317.

Ford, L. O., Johnson, L. N., Machin, P. A., Phillips, D. C. & Tjian, R. (1974). *Crystal structure of a lysozyme–tetrasaccharide lactone complex. J. Mol. Biol.* **88**, 349–371.

Furnas, T. C. (1957). *Single Crystal Orienter Instruction Manual*. General Electric Company, Milwaukee, USA.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *The structure of haemoglobin IV. Sign determination by the isomorphous replacement method. Proc. R. Soc. London Ser. A*, **225**, 287–307.

Green, D. W., North, A. C. T. & Aschaffenburg, R. (1956). *Crystallography of the β-lactoglobulins of cows' milk. Biochim. Biophys. Acta*, **21**, 583–585.

Hadfield, A. T., Harvey, D. J., Archer, D. B., MacKenzie, D. A., Jeenes, D. J., Radford, S. E., Lowe, G., Dobson, C. M. & Johnson, L. N. (1994). *Crystal structure of the mutant D52S hen egg white lysozyme with an oligosaccharide product. J. Mol. Biol.* **243**, 856–872.

Hamaguchi, K. & Imahori, K. (1964). *Optical measurements of the secondary structure of lysozyme. J. Biochem (Tokyo)*, **55**, 388–391.

Hamilton, W. C. (1955). *On the treatment of unobserved reflexions in the least-squares adjustment of crystal structures. Acta Cryst.* **8**, 185–186.

Hamilton, W. C., Rollett, J. S. & Sparks, R. A. (1965). *On the relative scaling of X-ray photographs. Acta Cryst.* **18**, 129–130.

Handoll, H. H. G. (1985). *Crystallographic Studies of Proteins*. DPhil thesis, University of Oxford.

Harker, D. (1956). *The determination of the phases of the structure factors of non-centrosymmetric crystals by the method of double isomorphous replacement. Acta Cryst.* **9**, 1–9.

Hart, R. G. (1961). *Refinement of heavy atom parameters other than relative y's. Acta Cryst.* **14**, 1194–1195.

Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C. & Rupley, J. A. (1972). *Vertebrate Lysozymes*. New York and London: Academic Press.

Jacobson, R. A., Wunderlich, J. A. & Lipscomb, W. N. (1961). *The crystal and molecular structure of cellobiose. Acta Cryst.* **14**, 598–607.

Jeffrey, G. A. (1990). *Crystallographic studies of carbohydrates. Acta Cryst.* B**46**, 89–103.

Johnson, L. N. (1966). *The crystal structure of N-acetyl-α-D-glucosamine. Acta Cryst.* **21**, 885–891.

Johnson, L. N. (1967). *An interaction between lysozyme and penicillin. Proc. R. Soc. London Ser. B*, **167**, 439–440.

Johnson, L. N. & Phillips, D. C. (1964). *Crystal structure of N-acetylglucosamine. Nature (London)*, **202**, 588–589.

Johnson, L. N. & Phillips, D. C. (1965). *Structure of some crystalline lysozyme–inhibitor complexes determined by X-ray analysis at 6 Å resolution. Nature (London)*, **206**, 761–763.

Jollès, J., Jaurrgui-Adell, J. & Jollès, P. (1964). *Amino-acid sequence & S–S bridges of HEWL. C. R. Acad. Sci. Paris*, **258**, 3926–3928.

Jolles, P. (1996). Editor. *Lysozymes: Model Enzymes in Biochemistry and Biology*. Basel, Boston, Berlin: Birkhauser Verlag.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature (London)*, **181**, 662–666.

Kendrew, J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C. & Shore, V. C. (1960). *Structure of myoglobin. A three-dimensional Fourier synthesis at 2 Å resolution. Nature (London)*, **185**, 422–427.

Kraut, J., Sieker, L. C., High, D. F. & Freer, S. T. (1962). *Electron density map of chymotrypsinogen at 6 Å resolution. Proc. Natl Acad. Sci. USA*, **48**, 1417.

Lavington, S. (1980). *Early British Computers: The Story of Vintage Computers and the People Who Built Them*. Manchester University Press.

Lemieux, R. U. & Huber, G. (1955). *Can. J. Res.* **33**, 128–133.

Lipson, H. & Cochran, W. (1968). *The Determination of Crystal Structures*. London: Bell.

Mo, F. & Jensen, L. H. (1975). *A refined model for N-acetyl-α-D-glucosamine. Acta Cryst.* B**31**, 2867–2873.

North, A. C. T. (1964). *Computer processing of automatic diffractometer data. J. Sci. Instrum.* **41**, 42–45.

North, A. C. T. (1965). *The combination of isomorphous replacement and anomalous scattering data in phase determination of non-centrosymmetric reflexions. Acta Cryst.* **18**, 212–216.

North, A. C. T., Phillips, D. C. & Mathews, F. S. (1968). *A semi-empirical method of absorption correction. Acta Cryst.* A**24**, 351–359.

Perutz, M. F. (1967). *Concluding remarks. Proc. R. Soc. London Ser. B*, **167**, 448.

Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H., Will, G. & North, A. C. T. (1960). *Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. Nature (London)*, **185**, 416–422.

Phillips, D. C. (1964). *On the design of single crystal diffractometers to measure a number of reflections simultaneously. J. Sci. Instrum.* **41**, 123–129.

Phillips, D. C. (1966). *The three-dimensional structure of an enzyme molecule. Sci. Am.* pp. 78–90.

Phillips, D. C. (1967). *The hen egg white lysozyme molecule. Proc. Natl Acad. Sci USA*, **57**, 484–495.

Poljak, R. J. (1963). *Heavy atom attachment to crystalline lysozyme. J. Mol. Biol.* **6**, 244–246.

Ramachandran, G. N., Ramakrishnan, G. & Sasisekharan, V. (1963). *Aspects of Protein Structure*. London: Academic Press.

Rollett, J. S. (1961). *Least-squares refinement in crystal-structure analysis*. In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, p. 122. London: Pergamon Press.

Rollett, J. S. & Sparks, R. A. (1960). *The correlation of intersecting layers of X-ray intensity data. Acta Cryst.* **13**, 273–274.

Rupley, J. A. (1964). *The hydrolysis of chitin by concentrated hydrochloric acid and the preparation of low molecular weight substrates for lysozyme. Biochim. Biophys. Acta*, **83**, 245–255.

Rupley, J. A. (1967). *The binding and cleavage by lysozyme of N-acetylglucosamine oligosaccharides. Proc. R. Soc. London Ser. B*, **167**, 416–428.

Scouloudi, H. (1960). *Structure of seal myoglobin in projection. Proc. R. Soc. London Ser. A*, **258**, 181.

Scouloudi, H. (1965). *The nature and configuration of the mercuri-iodide ion in the seal myoglobin derivative. J. Mol. Biol.* **12**, 17–26.

Stanford, R. H., Marsh, R. E. & Corey, R. B. (1962). *An X-ray investigation of lysozyme chloride crystals containing complex ions of niobium and tantalum: three-dimensional Fourier plot obtained from data extending to a minimum spacing of 5 Å. Nature (London)*, **196**, 1172–1177.

Steinrauf, L. K. (1959). *Preliminary X-ray data for some new crystalline forms of β-lactoglobulin and hen-egg-white lysozyme. Acta Cryst.* **12**, 77–79.

Stryer, L., Kendrew, J. C. & Watson, H. C. (1964). *The mode of attachment of the azide ion to sperm whale myoglobin. J. Mol. Biol.* **8**, 96–104.

Strynadka, N. C. J. & James, M. G. (1991). *Lysozyme revisited: crystallographic evidence for distortion of an N-acetyl muramic acid residue bound in site D. J. Mol. Biol.* **220**, 401–424.

Tipper, D. J. & Strominger, J. L. (1965). *Mechanism of action of penicillins: a proposal based on their structural similarity to acyl-D-alanyl-D-alanine. Proc. Natl Acad Sci. USA*, **54**, 1133–1141.

Vernon, C. A. (1967). *The mechanisms of hydrolysis of glycosides and their relevance to enzyme-catalysed reactions. Proc. R. Soc. London Ser. B*, **167**, 389–401.

Waser, J. (1951). *Lorentz factor for precession photographs. Rev. Sci. Instrum.* **22**, 563.

Watson, H. C., Kendrew, J. C., Coulter, C. L. & Brändën, C.-I. (1963). *Progress with the 1.4 Å resolution myoglobin-structure determination. Acta Cryst.* **16**, A81.

Wells, M. (1960). *Computation of absorption corrections on EDSAC II. Acta Cryst.* **13**, 722–736.

Wenzel, M., Lenk, H. P. & Schutte, E. (1962). *Z. Physiol. Chem.* **327**, 13–20.

Wilson, A. J. C. (1942). *Determination of absolute from relative X-ray intensity data. Nature (London)*, **150**, 151–152.

Wise, E. M. & Park, J. T. (1965). *Penicillin: its basis site of action as an inhibitor of a peptide cross-linking reaction in cell-wall mucopeptide synthesis. Proc. Natl Acad. Sci. USA*, **54**, 75–81.

Yang, J. T. & Doty, P. (1957). *ORD measurements on lysozyme. J. Am. Chem. Soc.* **79**, 761.

Subject index

A LA MODE, 663
A-DNA, 766
A-tract bending, 783
Ab initio phasing, 413
  in molecular replacement, 360
  low-resolution, 437
  multisolution methods, 417
Absolute configuration, 287, 372
Absorption, 60
Absorption coefficient
  atomic, 373
  atomic mass, 59
  linear, 59
Absorption corrections, 60
  for lysozyme, 860
Absorption edge, 54, 60
Absorption factor, 59
Accuracy, 499
  of unit-cell parameters, 266, 653
Acid–base equilibria, 733
ADIT, 827
ADIT-NMR, 827
Affinity chromatography, 87
Affinity tags, 92
Alignment of phase sets, 440
Alkylation of free cysteines, 96
All-atom contact analysis, 688, 690–692, 694–695, 697, 699
Alternative conformations, 492
American method of crystal orientation, 263
Amino acids
  distribution of water molecules around, 802
  hydrogen bonds in, 741
  interactions with heavy-atom reagents, 320
Amino-acid analogues as heavy-atom derivatives, 324
Amino-aromatic hydrogen bonding, 728
Ammonium sulfate, 86, 319
Amplitude contrast, 595, 599
Angular brightness of synchrotron radiation, 189
Angular reconstitution, 627
Angular refinement, 621
Animation, 448, 456
  of molecular-dynamics trajectories, 454
Anion-exchange chromatography, 86
Anisotropic atomic model, 486, 492
Anisotropic mosaicity, 300
Anisotropic scaling, 488
Annealing, 243
Annotation at the Protein Data Bank, 827
ANOLEA, 664
Anomalous correlation coefficient, 67
Anomalous Cullis *R* factor, 69
Anomalous difference Patterson map, 331
Anomalous phasing power, 69
Anomalous phasing with twinning, 548
Anomalous *R* factor, 67
Anomalous-scatterer labels for MAD, 377
Anomalous scattering (dispersion), 55, 369, 373
  data-collection strategies, 227
  heavy-atom location, 367, 371
  phase probability distribution, 370
  phasing, 367
  preparation of heavy-atom derivatives, 317
  signal quality, 298, 300
  sodium tartrate, 5
  without isomorphous replacement, 371
Anomalous scattering factors, 373
  evaluation of, 373
Anomalous substructure, 68
Antibody Fv fragments in crystallization of membrane proteins, 126
Antifreezes, 241
Antigen–antibody association, water molecules in, 815
Archimedes' method, 154
ARCIMBOLDO, 530
Area detectors, 178–179, 183, 266
  and cross fire, 163
  and synchrotron radiation, 195, 198
Argand diagram, 54
Aromatic halogen bonds, 824–825
ARP/wARP, 464, 525
Asn/Gln/His flips, 695
Astbury, W. T., 6
Astigmatism, 625
Asymmetric unit, 47
  choice of, 217
Atlas reports, 835
Atomic absorption coefficient, 373
Atomic charge distributions, 730
Atomic displacement parameters (temperature factors), 485
  anisotropic, 56, 485
  anisotropic, at atomic resolution, 492
  anisotropic, refinement in *SHELXL*, 531
  constraints, 490
  effect of coordinate errors, 461
  effect on coordinate uncertainty, 499
  group *B* factors, 461
  in structure validation, 656
  isotropic, 56
  refinement of, 461
  restraints, 490, 522
Atomic mass absorption coefficient, 59
Atomic radii, 708
  standard, 708
  van der Waals, 708, 713
Atomic resolution, 485
  anisotropic atomic displacement parameters, 492
  automatic location of water sites, 493
  data collection, 228
  data quality, 487
  definition of, 487
  deformation density, 494
  hydrogen atoms, 491
  ligands, 494
  low-resolution data, 488
  metal ions, 494
  modelling alternative conformations, 492
  ordered solvent, 493
  structures for validation, 673, 684
Atomic scattering factor, 54, 373
  imaginary component, 373
  real component, 373
Atomic solvation parameters, 715
Atomicity, 392, 486
AutoDep, 827
Autographa californica nuclear polyhedrosis virus (AcNPV), 83
Autoindexing, 263
  basis vectors, 264
  coordinate systems, 284
  distribution of reciprocal-lattice vectors, 263
  in *DENZO*, 283
  misindexing, 284
  twinning, 284
Autoinduction, 142
Automated convolution method for molecular-boundary identification, 387
Automated crystal mounting, 143
Automated model building, 525
Automated structure solution
  for MAD and MIR, 379
  in *PHENIX*, 544
Automatic location of water sites, 493
Automatic particle picking, 625
Automation in crystallography, 140
Automation of crystallization, 107
Automounters, 231
Averaging of reflection intensities, 298
  R factors, 298
AVGSYS, 361

B-DNA, 766
  A-tract bending, 783
  minor-groove width, 780
  sequence-dependent bendability, 781
  sequence-dependent deformability, 780, 783, 785
BABCOCK, 770
Babinet inverted structure, 360
Babinet's principle of complementarity, 499
Back-projection reconstruction, 621
Backbone geometry, 697
Backbone tracing, in *SHELXE*, 530
Background, determination of, 267
Background corrections, 60
Backrub, 691–692
Bacterial diseases, 16, 19
Bacteriophages, filamentous, 589
Bacteriorhodopsin, 123
  electron diffraction studies, 559–560
Baculovirus expression systems, 83
  AcNPV, 83
  BmNPV, 83
'Baking', 418
Balasubramanian plot, 655
Base pairing, 767
  Hoogsteen, 769
  Watson–Crick, 769
Basis vectors, 45
  in autoindexing, 264
  standard, 46
Batch methods of crystallization, 105
Baton building, 443
Beam divergence, 214
Beer's law, 59
Beevers, C. A., 6
Beevers–Lipson strips, 6
Bending magnets, 161
Bernal, J. D., 5–6
Bessel functions, 584
  spherical, 345
Best Fourier, 330, 369
Best phase, 330
Bias
  and restraints, 474
  avoidance, in *PHENIX*, 542
  in refinement, 405, 459
  model, 401, 404, 466, 472, 651
Biaxial crystals, 147
Bidentate hydrogen bonding, 756
Bifurcated hydrogen bonding, 721
Bijvoet, J. M., 5
Bijvoet differences, 375
Bijvoet pairs, 60, 374, 380
Bijvoet Patterson map, 9
BILDER, 449
Binary integer programming, 439
Binding energies, 715–716
  electrostatic contributions to, 734
Biological halogen bonds, 822
Biological Macromolecule Crystallization Database (BMCD), 838
BioMagResBank (BMRB), 831
Birefringence, 147
Blind region, 220, 273
Blindness, 23
Block-matrix approximation, 488, 505
Blow, D. M., 7, 9
Blow & Crick method, 368
Boltzmann model, 731
Bombyx mori nuclear polyhedrosis virus (BmNPV), 83
Bond-angle restraints, 474, 477, 479
Bond-length restraints, 474, 476–477
Bond lengths from the CSD, 738
Bovine α-lactalbumin, molecular-dynamics simulation, 645
Bovine pancreas ribonuclease A, molecular-dynamics simulation, 645
Bovine pancreatic trypsin inhibitor (BPTI)
  molecular-dynamics simulation, 633, 645
  solvent structure, 815
Bragg, W. H., 5
Bragg, W. L., 5–7, 845
Bragg's law, 5, 56, 65, 213
Bravais lattice, 52, 279
Bremsstrahlung, 159, 162
Britton plot, 314
5-Bromouracil, 822

5-Bromouridine, 378
Buckingham energy function, 587
Buffers, 319
  acetate, 319
  citrate, 319
  Tris buffer, 319
Bulk solvent, modelling of, 493
Bunn, C. W., 5
Buried water molecules, 726, 810

C^{α}-only model, 655
C^{β} deviation, 697
C41 strain, 93
C43 strain, 93
C—H···O hydrogen bonds, 727, 742
C—H···X hydrogen bonds, 741
Calcium-binding proteins, 757
  calmodulin, 757
  staphylococcal nuclease, 757
Calibration of CCD detectors, 184
Calmodulin, 757
Cambridge Structural Database (CSD), 736
  bond lengths, 738
  composite crystal-field environments, 742
  conformational libraries, 740
  conformations, 739
  data acquisition, 737
  data completeness, 737
  energies, 739
  hydrogen-bonding motifs, 743
  in structure validation, 662
  IsoStar, 743–744
  knowledge-based structural libraries, 738, 740
  metal coordination, 740
  relevance to protein crystallography, 736–737
  restraints from, 474
  software, 737
  SuperStar, 745
  van der Waals radii, 709
Cancers, 22
Carbohydrate-binding proteins, 755
Carbonyl–carbonyl interactions, 742
Carboxypeptidase, 9
Cardiovascular disorders, 23
Cartesian molecular dynamics, 469
CATH, 749–751
CCD detectors. *See* Charge-coupled device detectors
CCP4 (Collaborative Computational Project, Number 4), restraints, 662
Cell-free expression systems, 93
Cell lysis, 94
Cellular imaging, 234
Central limit theorem, 402–403
Centred (non-primitive) unit cell, 46
Centrosymmetric heavy-atom distributions, 371
Centrosymmetric point groups, 50
Channelling radiation, 159
Chaperones
  use in crystallization, 110
  use in protein folding, 81
Charge-coupled device (CCD) detectors, 183
  calibration, 184
  dark-current subtraction, 185
  flat-field corrections, 185
  for cryoTEM, 594, 599, 606
  geometric distortion, 185
  obliquity correction, 186
Charge distributions, atomic, 730
Charge-transfer bonds, 823
Charge-transfer equilibria, 733
Cheshire group, 349
Chiral volume, 654
Chirality, 145, 654
Chromatography, 86
  affinity, 87
  anion-exchange, 86
  dye-ligand, 86
  hydrophobic interaction, 86
  hydroxyapatite, 86
  immuno-affinity, 87
  size-exclusion, 86
Chymotrypsin, 9
CIF (crystallographic information file), 737
Circular variance plots, 656
Clashscore, 695
Class A metals, 318
Class B metals, 318, 321
Clotting factor Xa, 825
Cluster analysis, 739
  of phase sets, 441
CMC (critical micelle concentration), 94, 122, 125
Codon usage, effect on expression levels, 79
Coherent neutron scattering, 575
Coherent X-ray diffraction microscopy (CXDM), 234
Collaborative Computational Project, Number 4 (*CCP4*), restraints, 662
Combined molecular replacement, 349
Comparison of phase sets, 439
Complete rolling algorithm, 714
Complex formation, water molecules in, 815
Composite crystal-field environments, 742
Compton effect, 52
Compton scattering, 59
Concanavalin A, 502
Concentration of membrane proteins, 95
Conceptual clustering, 739
Conformational equilibria, NMR studies of, 618
Conformational heterogeneity, 131
Conjugate-gradient method, 397, 463, 488
  algorithms, 621
  comparison with simulated annealing, 471
  preconditioned, 463, 522–523
Connected rolling algorithm, 714
Connolly surface, 451, 714
Constraints, 461
  atomic displacement parameter, 490
  coordinate, 490
  for phase improvement, 396
  geometrical, 461
  in density modification, 386
  in *SHELXL*, 531
  nonlinear, 396
  real-space, 418
Contact surface, 707, 713
Contrast, 70
Contrast matching, 577, 580
Contrast-transfer function, 559, 596, 599, 606, 620
  correction, 621, 625
  in *EMAN*, 631
Contrast variation, 577, 580
Convex hull, 706
Cooperative hydrogen bonding, 721, 755, 758
Coordinate errors, 507
  effect on atomic displacement parameters, 461
  estimation of, 404, 657, 667
  estimation of, using *SFCHECK*, 667
  from cross validation, 466
  variable, probability distributions, 402
Coordinate systems, 284
Coordinate systems in *DENZO*
  beam–gravity, 284
  beam–spindle, 284
  beam–2θ, 285
  data, 284
Coordinate uncertainty, 499
  approximate methods, 505
  block-matrix calculations, 505
  low-resolution structures, 506
  modified Fourier method, 505
Coot, 443
CorA, 95
Corey–Pauling–Koltun (CPK) models, 450
γ correction, 395
  perturbation-γ approach, 395, 409
Correlation-coefficient translation function, 347
Correlation coefficients
  anomalous, 67
  CC(all), 68
  CC($F_{\rm obs}$, $F_{\rm calc}$), 71
  CC($I_{\rm obs}$, $I_{\rm calc}$), 71
  CC(weak), 68
  chain trace, 70
Coulombic potential, 731
Covariances, 499
Crambin
  deformation density, 494
  solvent structure, 814
Crick, F. H. C., 6–7
Critical angle for total external reflection, 165
Critical micelle concentration (CMC), 94, 122, 125
Critical wavelength, 161, 190
Cross-correlation function, 625
Cross fire, 162–163
Cross linking, 320
Cross-rotation function, 341
  and noncrystallographic symmetry, 335
Cross-translation function and noncrystallographic symmetry, 335
Cross validation, 464, 466, 491, 657
  in estimation of σ_A values, 405
  in maximum-likelihood refinement, 468
  restraints, 474
Cross vectors, 328–329
Crossed polarizers, 147
Crowther, R. A., 9
Crowther resolution, 237
Cryocrystallography, 241, 249, 256
  and MAD, 376
  antifreezes, 241
  apparatus, 244, 249
  cooling rates, 242
  cryogens, 242, 252
  cryoprotectants, 241, 249
  crystal mounting, 250, 252
  crystal storage, 253
  crystal transfer, 247, 253
  dual-stream apparatus, 245
  effect of crystal cooling on resolution, 243
  flash cooling, 252
  ice formation, 241–242, 249
  ice nucleation, 241
  preparation of crystals for, 249
  solvent modification, 241
  techniques, 249
  temperature calibration, 246
Cryo-electron microscopy, 620, 624
Cryo-electron tomography, 42–43
Cryogenic data collection, 231
Cryogens, 252
Cryoprotectants, 241, 249
  ethylene glycols, 241, 249
  glycerol, 249
  2-methylpentane-2,4-diol (MPD), 241, 249
  polyethylene glycol (PEG), 249
Cryotongs, 247
CRYSOL, 567
Crystal-density measurement, 152
  Archimedes' method, 154
  by flotation, 154
  by pycnometry, 154
  by tomographic crystal-volume measurement, 155
  by volumenometry, 154
  Ficoll density gradients, 155–156
  gradient-tube method, 155
  immersion microbalance, 154
Crystal engineering, 742
Crystal faces, 145
  indexing, 146
  properties of, 146
Crystal growth
  atomic force microscopy, 112
  conditions, 838
  diffusion, 102, 104–107
  electron microscopy, 112
  growth regime, 112
  habits, 145
  in convection-free media, 107
  interferometry, 112
  kinetics, 106, 110–112
  nucleation, 99, 103, 111, 115
  optical microscopy, 111–112
  phase diagrams, 99
  prediction of crystallizability, 107, 111, 114–115
  prenucleation, 111
  screw dislocation, 112
  two-dimensional islands, 112
Crystal habit, 145
  measurement of, 146
Crystal monochromators
  for neutrons, 170
  for X-rays, 166
  highly ordered pyrolytic graphite (HOPG), 166

- Crystal morphology, 145
- Crystal mounting, 145, 148
  - for cryocrystallography, 250, 252
  - loop mounting, 251
  - mechanical stability of, 243
- Crystal orientation
  - accuracy of, for data integration, 266
  - American method, 263
  - determination of, 263
- Crystal orientation matrix, 263
- Crystal quality, 135, 145
  - evaluation of, 108, 111, 113
  - mosaicity, 103, 113
  - perfection, 112, 115
- Crystal rotation, uneven, 290
- Crystal seeding, 110
- Crystal systems, 47, 52
  - hexagonal, 52
  - monoclinic, 47
  - orthorhombic, 47
  - tetragonal, 52
  - triclinic, 47
  - trigonal, 52
- Crystal-to-detector distance, choice of, 221
- Crystallants, 100
- Crystallization, 838
  - automation of, 107
  - batch methods, 105
  - by water evaporation, 106
  - counter-diffusion method, 107
  - dialysis methods, 105
  - dynamic light scattering, 104, 111
  - effect of electric field, 102, 107, 110
  - effect of magnetic field, 107, 110
  - effect of pH, 101, 106
  - effect of pressure, 105, 107, 110
  - effect of temperature, 101, 105–106, 110
  - fluorescence spectroscopy, 111
  - hanging-drop method, 106
  - high-throughput method, 105, 109–110
  - in gels, 108
  - in microfluidic devices, 109
  - interface diffusion method, 104, 107
  - ionic liquids, 101, 103
  - of lysozyme, 101, 103, 108, 110–111, 113, 846
  - of membrane proteins, 122–123
  - phase diagrams, 99, 114
  - precipitants, 123
  - promotion of, 140
  - screening, 113
  - sitting-drop method, 106
  - solubilities, 99, 101, 114–115
  - supersaturation, 99
- Crystallization additives, 102, 123
- Crystallization database, 114
- Crystallization chaperones, 110
  - noncovalent, 132
- Crystallization robots, 142
- Crystallization screens, 840
- Crystallization strategies, 838
- Crystallizing agents, 100
- Crystallographic competition assay, 824
- Crystallographic information file (CIF), 737
- Crystallographic *R* factor, 71
- *Crystallography & NMR System* (*CNS*), 463, 472
- CSD. *See* Cambridge Structural Database
- Cubic bicontinuous lipidic phases, 126
- Cullis *R* factor, 69
- Curved single-crystal monochromators, 196
- *CURVES*, 770
- Cyclooxygenase 1, 124
- Cyclooxygenase 2, 124
- Cylindrical averaging in fibre diffraction, 583
- Cylindrical coordinates, 584
- Cylindrically averaged Patterson function, 586
- Cytochrome bc_1 complex, 123
- Cytochrome *c*, 9
- Cytochrome *c* oxidase, 123
  - crystallization of, 126

- $D^{\ell}_{m,m'}$ matrices, 345
- $D_2O - H_2O$ difference maps, 554–555, 801
- DALI, 749–751
- DALI domain dictionary, 751, 753
- *Daresbury Laue Software Suite*, 207
- Dark-current subtraction for CCD detectors, 185
- DARPin chaperones, 133
- Data collection, 65
  - exposure time, 224
  - fine slicing, 215
  - geometries, 212
  - in electron diffraction, 557
  - in fibre diffraction, 585
  - in SAXS, 570
  - low-temperature, 241, 249
  - monochromatic, 211
  - rotation method, 213
  - rotation range, 215–217, 219
  - still exposure, 213
  - strategies, 226
  - wide slicing, 215
- Data completeness, 66, 211
  - and direct methods, 414
  - and refinement, 460
  - and structure validation, 653
  - in the rotation method, 217
- Data integration, 266
  - accuracy, 266
  - background determination, 267
  - by profile fitting, 266, 268
  - by summation, 266–267
  - for fibre diffraction, 586
  - standard profiles, 268
- Data processing
  - autoindexing, 263
  - for lysozyme, 850, 859
  - in electron diffraction, 558
  - in fibre diffraction, 585
  - in SAXS, 571
  - integration, 266
  - partially recorded reflections, 296
  - *DENZO* and *SCALEPACK*, 282
- Data quality, 65
- Data redundancy, 206–207, 217, 227, 376, 653
- Data resolution
  - and direct methods, 413
  - and refinement, 460
  - and structure validation, 653
  - effective, 653
  - nominal, 653
- Data-to-parameter ratio in fibre diffraction, 587
- Databases
  - Cambridge Structural Database, 736
  - for crystallization, 114
  - heavy-atom data bank, 317, 325
  - Nucleic Acid Database, 833
  - Protein Data Bank, 827
- Debye equation, 567, 575
- Debye function, 565
- Debye–Scherrer arcs, 584
- Decay *R* factor, 68
- 3Dee, 749–750
- Defocus series, 620
- Deformation density, 494
  - in crambin, 494
- Delaunay triangulation, 705
- *DEMON/ANGEL*, 361
- Density modification, 385
  - constraints, 386
  - in *DM/DMMULTI*, 407
  - in *PHENIX*, 544
  - quality indicators for, 69
  - *R* factors, 70
  - reciprocal-space interpretation of, 393
  - scaling of observed structure factors, 390
- Density-modification free *R* factor, 70
- Density-modification real-space free residuals, 70
- *DENZO*, 263, 269, 282
- De-orthogonalization matrix, 355
- Deposition sites for the Protein Data Bank, 827
- Detective quantum efficiency (DQE), 177, 183
- Detector calibration, 184, 266
- Detector distortions, 289
- Detector overloads, 224, 271
- Detectors
  - accuracy, 177
  - area, 178–179, 183, 266
  - CCD, 183
- Detectors
  - dynamic range, 178
  - film, 180
  - flat-field corrections, 178
  - gas discharge (wire) counters, 180
  - geometric distortion, 178
  - image plates, 172, 180
  - line spread function, 177
  - multiwire proportional counters, 171, 180
  - neutron, 171
  - photon counters, 179
  - photon integrators, 179–180
  - pixel array, 181
  - point, 179
  - point spread function, 177
  - reverse-biased semiconductor, 180
  - scintillator/photomultiplier, 179
  - signal-to-noise ratio, 177
  - spatial resolution, 177
  - stopping power, 178
  - storage phosphors, 180
  - synchrotron-radiation, 195
  - television, 181
  - two-dimensional, 212
  - X-ray, 177, 183
- Detergents, 122, 125
  - as solubilization agents, 103
  - for membrane-protein solubilization, 94
  - in membrane-protein crystallization, 123
- Detwinning, 548
  - model-based, 550
- Diabetes, 23
- Diagonal-approximation method, 397, 488
- Dialysis methods of crystallization, 105
- Dickerson, R. E., 7
- Dielectric constant, 730
  - effective, 732
- Difference density quality, 658
- Difference distance matrix plot, 454
- Difference Fourier maps, 329
  - in fibre diffraction, 588
- Difference Fourier syntheses for heavy-atom location, 371
- Differential phase residual, 601
- Diffraction-component precision index (DPI), 73, 506–507, 658, 667
  - examples, 507
- Diffraction-pattern prediction, in *DENZO*, 288
- Diffraction-pattern symmetry, 60
- Diffraction physics, 52, 282
- Diffraction ripples in Patterson maps, 62
- Diphtheria toxin, structure validation, 681
- Direct methods, 68, 413
  - and maximum entropy, 433
  - false minima, 422
  - limitations, 435
  - minimal function, 417
  - multisolution methods, 417
  - parameter-shift method, 418
  - peak picking, 418
  - peaklist optimization, 418
  - tangent formula, 417
- Directed evolution, 130
- Directional atomic contact analysis, 655
- Disorder, 462
  - dynamic, 462, 485
  - modelling in *SHELXL*, 532
  - rotational, 584
  - screw, 584
  - static, 462, 485
  - translational, 584
- Displaying information, 454
- Displaying structures, 448–449
  - animation, 456
  - CPK models, 450
  - direct volume rendering, 453
  - geometric representation, 450
  - illustration, 455
  - isosurfaces, 452
  - physical models, 456
  - ray tracing, 451
  - stereolithography, 456
  - three-dimensional printing, 456
  - volumetric representation, 452

Distance matrix plot, 454
Distance measurements, 579–580
DM/DMMULTI, 361, 407
DNA, 6, 766, 778
  A, B and Z helices, 777
  A-DNA, 766
  B-DNA, 766
  B-to-Z interconversion, 778
  handedness of helices, 774
  hydrogen bonding in, 726
  interactions with proteins, 757, 759
  major groove, depth of, 774
  minor groove, depth of, 774
  Nucleic Acid Database, 833
  torsion angles, 771
  treatment in *PHENIX*, 541
  Watson–Crick Z-DNA, 778
  Z-DNA, 766
DNA/RNA hybrids, 766
DOCK, 745
Dodecylphosphatidylcholine, 96
Double-crystal monochromators, 196
DPI. *See* Diffraction-component precision index
DPS, 265
DQE (detective quantum efficiency), 177, 183
DREAR, 415
Drug delivery, 41
Drug design, 15, 135
  use of the CSD, 745
Drug metabolism, 25
Drug resistance, 18
Dual-axis tomography, 43
Dual-space methods, 68
Duplex RNA, 766
Dye-ligand chromatography, 86
Dynamic disorder, 462, 485
Dynamic processes
  NMR studies of, 615, 618
  SAXS studies of, 563

E. coli expression systems, 77, 79
Early structure determinations, 6, 8
  carboxypeptidase, 9
  chymotrypsin, 9
  cytochrome *c*, 9
  globular proteins, 6–7
  glyceraldehyde-3-phosphate dehydrogenase, 10
  haemoglobin, 6–8
  insulin, 6, 9
  lactate dehydrogenase, 9
  lysozyme, 9, 845
  membrane proteins, 9
  myoglobin, 6, 8
  papain, 9
  penicillin, 5
  pepsin, 6
  ribonuclease, 9
  sodium tartrate, 5
  tobacco mosaic virus, 6
  viruses, 9
  vitamin B$_{12}$, 5
EDSAC1 and 2, 8
Effective dielectric constant, 732
Effective resolution, 66
Elastase, solvent structure, 809
Elastic scattering, 52
Electron cryomicroscopy (cryoTEM), 593
  Protein Data Bank, 827
  specimen preparation, 597
  types of specimens, 597
  vitrification of specimens, 597
Electron crystallography, 557
  of membrane proteins, 561
Electron density
  calculation of, 60
  skewness of, 70
Electron-density averaging, 352, 390–391
  computer programs, 361
  convergence, 359
  in *DM/DMMULTI*, 409, 411
  multidomain, 360
  multiple-crystal-form, 358, 360
Electron diffraction, 557
  bacteriorhodopsin, 559–560
  data collection, 557
  data processing, 558
  data sampling, 558
  radiation damage, 558
  refinement, 560
  specimen preparation, 557
  structure factors, 559
  tubulin, 560–561
Electron microscopy, 557, 593
  amplitude contrast, 599
  contrast-transfer function, 559, 599
  electron sources, 557
  field emission gun, 595
  impact on structural biology, 42
  minimal (low-dose) procedure, 558
  phase contrast, 599
  recording media, 557
  signal-to-noise ratio, 596
  spatial coherence of electron beam, 595
  spherical aberration, 599
Electron pair distribution function, 565–566
Electron scattering, 557, 593
  elastic, 593–594
  inelastic, 593–594
Electron tomography, 620
Electron-transfer equilibria, 733–734
Electrostatic energy, 732
Electrostatic fields, response of proteins to, 730
Electrostatic force, 732
Electrostatic free energy, 732
Electrostatic interactions in proteins, 730
Electrostatic polarization, 730
Electrostatic potential
  calculation of, 732
  distributions, 730, 732
Electrostatics, 535, 730
  acid–base equilibria, 733
  Boltzmann model, 731
  Coulombic potential, 731
  dielectric constant, 730
  effective dielectric constant, 732
  electron-transfer equilibria, 733–734
  electrostatic polarization, 730
  ion binding, 733–734
  Langevin model, 730–731
  linear solvent dielectric models, 731
  point inducible dipole, 730
  Poisson–Boltzmann equation, 731
  polarity, 731
  polarizability, 730–731
  reaction potential, 731
  screening potential, 731
  self-potential, 731
  solvation potential, 731
EMAN, 629
Enantiomorph, 360
  discrimination with *SHELXE*, 530
  selection of correct, 377
Enantiomorphic point groups, 47–48
Enantiomorphism, 46
Ensemble, 433
ENTREZ, 749, 751
Entropy, 433
  Shannon's theorems, 433
ENVelope, 361
Environment profiles, 663
Enzyme catalysis, role of metal ions, 756
Epitaxial twinning, 311
ERRAT, 678
Error model, 67
Error-reduction algorithm, 234
Errors
  coordinate, 507
  coordinate, estimation of, 404, 657, 667
  coordinate, from cross validation, 466
  coordinate, probability distributions for, 402
  detector, 268
  effect on MAD phasing, 376
  in protein-structure models, 520, 649, 651–652, 677
  instrument, 268
  position, 507
  systematic, in profile-fitted intensities, 271
Estimated standard deviation (e.s.d.), 499, 501
Ethylene glycols, 241, 249
Euler parameterization, 345
Ewald, P. P., 5
Ewald sphere, 57–58, 205, 213, 272
  partial reflections, 296, 301–302
Expected intensity factor, 402
Exposure, uneven, 290
Exposure time, 224
Expression systems, 76–77
  baculoviruses, 83
  constitutive, 78
  constructs, 76
  E. coli, 77, 79
  fermentation, 80
  growth media, 80
  inducible, 78
  insect cell–virus, 83
  mammalian cells, 84
  misfolded proteins, 80
  plasmids, 78
  post-translational modifications, 75–76, 82, 89
  preparation of cDNA clones, 76
  T7 polymerase, 78
  yeasts, 82
Extended-atom parameters, 642
Extended atoms, 715
Extinction, 59

^{19}F NMR, 821
F$_1$ ATPase, synchrotron-radiation studies of, 198
Fab antibody fragments, 95
False minima, 422
Fankuchen, I., 6
Fast Fourier transform (FFT), 489
  in *TNT*, 522
Fast rotation function, 342
  angular resolution, 343
Fast screen, 838
FepA, 124
Few-atoms model, 439
FhuA, 124
Fibre diffraction, 583–584
  background subtraction, 586
  cylindrical averaging, 583
  data collection, 585
  data integration, 586
  data processing, 585
  Debye–Scherrer arcs, 584
  difference Fourier maps, 588
  early studies, 6
  large assemblies, 590
  layer lines, 584
  polynucleotides, 589
  polypeptides, 588
  polysaccharides, 589
  profile fitting, 586
  refinement, 587
  rotational disorder, 584
  screw disorder, 584
  selection rule, 584
  simulated annealing, 588
  structure determination, 586
  structure factors, 584
  time-resolved studies, 585
  tobacco mosaic virus, 587
  translational disorder, 584
Fibre-optic tapers, 183
Fibres, 583
  noncrystalline, 583
  polycrystalline, 583
  preparation of, 585
Ficoll density gradients, 155–156
Figure-of-merit weighting for model phases, 404
Figures of merit, 69–70
Filamentous bacteriophages, 589
Fine slicing, 215
Flash cooling, 252
Flat-field corrections for CCD detectors, 185
FLExS, 745
FLExX, 745
Flood-field detector calibration, 287
Flotation method, 154
Flux of synchrotron radiation, 189

- Focusing collimators
  - capillary optics, 165
  - for microfocus sources, 165
- Force constants, 475
- Force-feedback devices, 457
- Force fields, 644
  - *CHARMM*, 662
  - *CHARMM22*, 644
  - *GROMOS96*, 633–634, 640
  - modifications for structure determination, 644
  - parameterization of, 644
- Fourier maps, difference, 329
- Fourier methods, 5
- Fourier refinement, 419
- Fourier shell correlation, 43, 601–602, 621
- Fourier summation, 62
- Fourier syntheses, histogram of, 438
- Fourier transformation, 60
  - in *DM/DMMULTI*, 411
  - inverse, 60
- Fourier–Bessel structure factor, 584
- Fourier–Bessel syntheses, 584
- Fractional atomic coordinates, 60
- Frame shift, 649
- Franklin, R., 7
- Free-atom model, 525
- Free-electron model, 373
- Free lunch algorithm, in *SHELXE*, 530
- Free phase residual, 561
- Free *R* factor, 71, 466, 506, 651, 657, 665
  - for density modification, 70
  - precision of, 665
- *FREEHELIX*, 770
- Fresnel zone, 54
- Friedel pairs, 60, 375
- *FRODO*, 449
- Frost prevention, 245
- Full-matrix inversion
  - restrained, 502
  - unrestrained, 502, 504
- Fully recorded reflections, 214
- Fungi, 17, 21
- Fusion chaperone, 132
- Fusion proteins, 77, 131, 140
- Future of crystallography, 39
- Future of science, 39

- *G* factors, 656
- *GAP*, 361
- Gauss–Bonnet theorem, 715
- Geis, I., 766, 871
- Gelder's law, 449
- Gels, crystallization in, 108
- General purpose graphics processing unit, in *EMAN*, 630
- Generation of phase sets, 438
- Genetic algorithms, 330, 567
- Genetic diseases, 14
- Geometric distortion in CCD detectors, 185
- *Glide*, 534, 537
- Global minimum, 466
- Glutaraldehyde, 320
- Glyceraldehyde-3-phosphate dehydrogenase, 10
- Glycerol, 249
- Glycosyl bond geometry, 773–774
  - *anti*, 774
- Glycosylation, 75, 82–83, 133
- *GOLD*, 745
- Goodness of fit, 501
- Gouraud shading, 451
- Gradient-tube method, 155
- *GRAMPS/GRANNY*, 449
- *GRASP*, 450, 714, 718
- *GRID*, 745
- *GRIP*, 449
- Gripper, 232
- *GROMOS96*, 633–634, 640
- GST tag, 92
- Guinier plot, 565, 577

- *h*-cell, 354
- HAD (heavy-atom data bank), 317, 325
- Haemoglobin, 6–8
  - dimeric, water molecules in, 816
- α-Haemolysin, 124
- Halogen-bond acceptors, 823–824
- Halogen-bond geometry, 823
- Halogen bonds *versus* hydrogen bonds, 824
- Halogen interactions, 821
  - in the Protein Data Bank, 822
- Halogens, *in vivo* incorporation of, 821
- Hamilton, Rollett and Sparks method, 296
  - for partial reflections, 296
  - singular-value decomposition of the normal-equations matrix, 296
- Hamilton's significance test, 588
- Hammett σ constants, 821–822
- Handedness, 145, 329, 360, 371, 382
- Hanging-drop method, 106
- Hard cations, 321
- Hard ligands, 318
- Hard metals, 756
- Harker, D., 7
- Harker lines, 63
- Harker phase diagram, 368
- Harker planes, 63
- Harker sections, 328
- Hassel, O., 822
- *HBPLUS*, 722
- *HEAVY*, 330
- Heavy-atom data bank (HAD), 317, 325
- Heavy-atom derivatives
  - amino-acid analogues, 324
  - of lysozyme, 846, 855
  - preparation of, 317
- Heavy-atom distributions, centrosymmetric, 371
- Heavy-atom location, 327, 367, 371
  - and direct methods, 425
- Heavy-atom reagents, 318
  - class B metals, 321
  - effect of concentration, 320
  - effect of pH, 318–319
  - effect of precipitants and buffers, 319
  - effect of temperature, 320
  - effect of time of soak, 320
  - electrostatic binding of, 323
  - hard cations, 321
  - hydrophobic, 323
  - interactions with amino acids, 320
  - lability, 318
  - oxidation states, 318
  - polynuclear, 324
  - solubility, 319
  - stability, 318
- Helical particles, 603
- Helical symmetry, 333, 584
- Helical viruses, 589
- α-Helices, 6
- Helices, hydrogen bonding in, 723
- Helix capping, 725
- Helix parameters
  - *BABCOCK*, 770
  - *CURVES*, 770
  - *FREEHELIX*, 770
  - horizontal displacement, 770
  - in nucleic acids, 766, 770
  - inclination, 770
  - *NEWHELIX*, 770
  - propeller, 770, 775
  - rise, 775
  - rise per base pair, 770
  - roll, 770
  - twist, 770, 775
  - x displacement, 770
- Helix-termination motifs, 724
- Helminths, 18, 21
- Hemihedral twinning, 146, 312
- Hendrickson–Lattman coefficients, 394, 407
- Hetero groups, structure validation, 663
- Hexagonal crystal system, 52
- High-throughput cloning, 140–141
- High-throughput crystallization, 142
- High-throughput crystallography, 140
- High-throughput screening methods, 134
- Highly ordered pyrolytic graphite (HOPG), 166
- Histogram of a Fourier synthesis, 438
- Histogram matching, 388
  - in *DM/DMMULTI*, 407–408, 411
- *HKL*, 282
- Hodgkin, D. C., 5
- Holliday junction, 822
- Holmes, K. C., 7
- HOMSTRAD, 749–750
- Hoogsteen base pairing, 588, 769
- Hoppe, W., 10
- Huber, R., 10
- Hybrid model, 525
- *HYDRA*, 449
- Hydration, estimation of, 153
- Hydration surface, 714
- Hydrogen atoms
  - at atomic resolution, 491
  - in *PrimeX*, 535
  - neutron diffraction, 553
- Hydrogen-bond acceptors, 824
- Hydrogen bonding, 721
  - amino-aromatic, 728
  - analysis of in structure validation, 651, 655
  - and secondary structure, 723, 742
  - bidentate, 756
  - bifurcated, 721
  - C—H⋯O, 727, 742
  - C—H⋯*X*, 741
  - cooperative, 721, 755, 758
  - *HBPLUS*, 722
  - in amino acids, 741
  - in helices, 723
  - in nucleic acids, 722, 726
  - in *PrimeX*, 535
  - in proteins, 721, 723
  - in β-sheets, 724
  - in side chains, 724
  - in turns, 724
  - in ubiquitin, simulation of, 635–637
  - involving sulfur, 727
  - local, 724
  - N—H⋯carbonyl, 741
  - N—H⋯π, 742
  - O—H⋯O, 741
  - O—H⋯π, 742
  - protein–water, 725
  - resonance-assisted, 741
  - resonance-induced, 741
  - short, in phosphate-binding protein, 763
  - transition metals as proton acceptors, 742
  - use of the CSD, 740
- Hydrogen-bonding criteria, 722
- Hydrogen-bonding motifs, 725, 743
- Hydrogen-bonding patterns, 721
- Hydrogen-bonding potential, 721–722
  - saturation of, 723
- Hydrogen/deuterium exchange, 556
- Hydrophobic effects in halogen interactions, 821
- Hydrophobic interaction chromatography, 86
- Hydrophobicity, 714–715
- Hydroxyapatite chromatography, 86
- Hyperglycosylation, 82

- $I/\sigma(I)$ ratio, 224–225, 228, 653
- Ice formation, 241–242, 249
  - nucleation, 241
  - prevention of, 241
- Icosahedral particles, 603
- Icosahedral point groups, 52
- Icosahedral symmetry, 47
- Image plates
  - for neutrons, 172
  - for X-rays, 180
- Image reconstruction, 620
- *IMAGIC*, 630
- *IMAGIC 4D*, 624
- Imaging of whole cells, 234
- Immersion microbalance, 154
- Immobilized metal affinity chromatography (IMAC), 95
- Immuno-affinity chromatography, 87
- Implicit solvation, 535
- Incoherent neutron scattering, 575
- Indexing, 274
  - alternative schemes, 221
  - autoindexing, 263, 283
  - basis vectors, 264

Indexing
  distribution of reciprocal-lattice vectors, 263
  of crystal faces, 146
Inelastic scattering, 52
Infectious diseases, 16
Information content of crystallographic data, 487–488
Information measure, 487
  quadratic, 487
Input–output algorithm, 234
Insect cell–virus expression systems, 83
Insertion devices, 161, 190
  multipole wigglers, 190
  periodic magnet, 191
  undulators, 190
  wavelength shifters, 190
Insertion fusion, 132
Insight II, 718
Instrument errors, 67
Insulin, 6, 9
Integration of diffraction data, 266, 275
  accuracy, 266
  background determination, 267
  by profile fitting, 266, 268
  by summation, 266–267
  from fibres, 586
  standard profiles, 268
Inteins, 77
Intensity-based likelihood refinement, 489
Intermolecular interactions
  data from the CSD, 740
  energies of, 45
  weak, 742
Intermolecular perturbation theory, 742
Internal contrast, variation of, 577
International Tables for Crystallography, 5
Internationale Tabellen zur Bestimmung von Kristallstrukturen, 5
Inverse-beam geometry, 375
Inverse Fourier transformation, 60
Inversion symmetry, 46
Ion binding, 733–734
Isoelectric focusing, 88
Isomorphism, lack of, 330
Isomorphous replacement, 7, 367
  and fibre diffraction, 586
  data-collection strategies, 226
  height of peaks in a Patterson map, 62
  location of heavy-atom sites, 327
  noncrystallographic symmetry, 331
  preparation of heavy-atom derivatives, 317
  pseudosymmetry, 331
IsoStar, 743–744
Isosurface representations, 452
Isotopic substitution
  statistical labelling method, 580
  triple, 578
Iterative model building, 526

Jaynes' maximum-entropy formalism, 434
Jaynes' maximum-entropy principle, 434
JIFFILOOP, 699

$K\alpha$ edge positions of different elements, 54
KcsA, 95
Kendrew, J. C., 7
Kendrew models, 10
Ketosteroid isomerase, 825
Kinemages, 688, 694, 699
KiNG, 688, 694, 698–699
Klug, A., 7
Knowledge-based interaction potentials, 663
Kramers–Kronig transform, 55

L_1 norm, 460
L_2 norm, 460
Lack of closure error, 329, 370
Lack of isomorphism, 330
Lactate dehydrogenase, 9
LALS (linked-atom least squares), 587
Langevin model, 730–731
Lattice plane, 45
Lattice point, 45
Lattice pseudosymmetry in autoindexing, 284
Lattice symmetry in autoindexing, 283
Lattice-translocation disorder, 315
Laue, M. von, 5
Laue conditions, 56
Laue diffraction, 205
  radiation damage, 207
  time-resolved studies, 205, 208
  use of synchrotron radiation, 195, 205
  wavelength normalization curve, 195, 206
Laue equations, 272
LaueView, 207
Layer-line splitting, 587
Layer lines in fibre diffraction, 584
LEAP, 207
Least-squares full matrix, variances and covariances from, 499
Least-squares methods, 460, 488, 500
  in *SHELXL*, 531
  normal equations, 500
Legume lectins, solvent structure, 808
Length distribution, 576
Lennard–Jones 6–12 potential, 642
Lewis acids, 823
Lewis bases, 823
Ligand fitting
  in *Coot*, 444
  in *PHENIX*, 541
  in *PrimeX*, 537
Ligands at atomic resolution, 494
Ligation-independent cloning, 141
Light-harvesting complex 2, 123
Light-harvesting complex II, 123
Likelihood-based selection of masks, 439
Linderström-Lang, K. U., 8
Linear absorption coefficient, 59
Linear diffractometer, 848, 856
Linear solvent dielectric models, 731
Linked-atom least squares (*LALS*), 587
Lipids, 96
Lipson, H., 6
Liquid helium, 244
LLG-Z score, 71
Local hydrogen bonds, 724
Local intensity difference, 313
Local minima, 466
Local scaling, 381
  in *TNT*, 522
Locked rotation function, 343
Locked translation function, 350
Log-likelihood gain, 71
Lonsdale, K., 5–6
Loop refinement, in *PrimeX*, 537
Lorentz factor, 59–60, 273
  errors in, 285
Low-abundance tRNAs, 79
Low-resolution *ab initio* phasing, 437
Low-resolution data
  importance of, 228
  in atomic resolution refinement, 488
Low-resolution images, 437
Low-resolution structures, coordinate uncertainty, 506
LUDI, 745
Lunes, 214
Luzzati distribution, 403
Luzzati plot, 73, 404, 509, 658
Lysozyme, 9, 845
  absorption corrections, 860
  at 2 Å resolution, 854, 863
  at 6 Å resolution, 845, 852–853
  biological function of, 866
  calculation of phase values, 862
  catalytic mechanism of, 868
  crystal-type problem, 861
  crystallization, 846
  data processing, 850, 859
  heavy-atom derivatives, 846, 855
  hen egg-white, molecular-dynamics simulation, 642, 645
  intensity measurements at high resolution, 858
  model in crystallization studies, 111, 113
  refinement of heavy-atom parameters, 862
  scaling intensity data, 851
  T4, mutants, 824
  T4, solvent structure, 812
Machine emittance, 190
Machine learning, 739
Macromolecular crystallographic information file (mmCIF), 663, 737, 833
MAD. See Multiwavelength anomalous diffraction
Maestro, 534, 537
MAGICSQUASH, 361
Magnesium sulfate, 319
Maltoporin, 124
Mammalian-cell expression systems, 84
Mammalian-cell inducible promoters, 85
Map connectivity, 438
Marching-cube algorithm, 453, 714
Masks, likelihood-based selection, 439
Mass spectrometry, 88
Matthews number, 152
MAVE, 361
Maximum entropy, 433
  and crystallography, 435
  equations, 434
  Jaynes' formalism, 434
  Jaynes' principle, 434
Maximum likelihood, 460, 489
  estimation of phase errors, 403
  in *CNS*, 517
  structure refinement, 467
Maximum-likelihood minimization, in *PrimeX*, 536
MDIR (multidimensional isomorphous replacement) in fibre diffraction, 587
Mean anomalous signal-to-noise ratio, 68
Mean signal-to-noise ratio, 67
Mechanosensitive ion channel, 123
Medicine and crystallography, 13
  bacterial diseases, 16, 19
  blindness, 23
  cancers, 22
  cardiovascular disorders, 23
  diabetes, 23
  fungi, 17, 21
  genetic diseases, 14
  helminths, 18, 21
  infectious diseases, 16
  neurological disorders, 23
  protozoan infections, 16, 21
  structure-based drug design, 15
  viruses, 16–17
Membrane proteins, 123
  bacteriorhodopsin, 123
  crystallization, 122–123
  crystallization using additives, 123
  crystallization using antibody Fv fragments, 126
  crystallization using cubic bicontinuous lipidic phases, 126
  crystallization using detergents, 123, 125
  cyclooxygenase 1, 124
  cyclooxygenase 2, 124
  cytochrome bc_1 complex, 123
  cytochrome *c* oxidase, 123, 126
  electron crystallography of, 561
  expression and purification of, 92
  FepA, 124
  FhuA, 124
  α-haemolysin, 124
  light-harvesting complex 2, 123
  light-harvesting complex II, 123
  maltoporin, 124
  mechanosensitive ion channel, 123
  OmpA, 124
  OmpF, 124
  PhoE, 124
  photosynthetic reaction centres, 9, 123
  porins, 124
  potassium channel, 123
  precipitants, 123
  prostaglandin H_2 synthase, 124
  squalene cyclase, 124
Merging *R* factors, 653
  R_{meas}, 66, 653
  R_{merge}, 66, 653
  $R_{\text{p.i.m.}}$, 66
  $R_{\text{r.i.m.}}$, 66
Merohedral twinning, 311
Metal coordination geometry, use of the CSD, 740
Metal ions at atomic resolution, 494

Metalloproteins, metal-ion replacement in, 324
Metcalf's law, 449
Methane monooxygenase hydrolase, 821
Methyl-group conformations from neutron diffraction, 554
2-Methylpentane-2,4-diol (MPD), 241, 249
Metropolis Monte Carlo simulation, 469
Micelles, 94–96, 122
Microfluidics, 109
Microfocus X-ray tubes, 160
Microgravity, 107, 113, 838
  effect on crystallization, 102
  gravity and hypergravity, 110
  mimicking of, 107, 109–110
Microspectrophotometry, 259
MIDAS, 449
Miller indices, reduced, 296
Minimal function, 69, 417
Minor-groove-binding drugs, 790
MIR. *See* Multiple isomorphous replacement
Mirror symmetry, 46
Misfolded proteins, 80
Misindexing, 284
Mis-setting angles, 297
Missing symmetry, 654
MLF target function, 468
MLHL target function, 468
MLI target function, 468
mmCIF (macromolecular crystallographic information file), 663, 737, 833
MMSX, 449
Model bias, 401, 466, 472, 651
  in combined phase maps, 404
  in figure-of-merit weighted maps, 404
Model building
  in *Coot*, 443
  in *PHENIX*, 541, 544
Model evaluation, 464
Model/noise bias, 632
Model phases, figure-of-merit weighting, 404
Model rebuilding, 651
Moderators, 169
  for cold neutrons, 169
  for spallation sources, 173
  for thermal neutrons, 169
Modified Fourier method for estimating coordinate uncertainty, 505
Modified Patterson functions, 329
Molecular biology, 75
Molecular-boundary identification by automated convolution method, 387
Molecular design, 39
Molecular dynamics, 633, 642
  animation of trajectories, 454
  average structures, 644
  Cartesian, 469
  effect of crystallographic resolution, 645
  extended-atom parameters, 642
  in fibre diffraction, 588
  in simulated annealing, 463, 469
  internal motions, 644
  particle mesh Ewald method, 643
  potential-energy functions, 642
  relaxation periods, 634
  restraints, 633, 643
  simple point charge model, 633–634
  torsion-angle, 469–470
  united-atom approach, 633
Molecular-dynamics simulation
  bovine pancreas ribonuclease A, 645
  bovine α-lactalbumin, 645
  BPTI, 633, 645
  hen egg-white lysozyme, 642, 645
  trypsin, 645
  ubiquitin, 633
  ubiquitin, atomic mean-square position fluctuations, 636
  ubiquitin, averaging period, 638
  ubiquitin, dihedral-angle fluctuations, 639
  ubiquitin, hydrogen bonding, 635–637
  ubiquitin, internal motions, 639
  ubiquitin, potential energy, 634
  ubiquitin, root-mean-square atom-position deviation, 634
Molecular-dynamics simulation
  ubiquitin, translational and rotational fitting, 636
  ubiquitin, water diffusion, 640
Molecular envelopes, 352
  determination of, 356
Molecular graphics, 448–449, 534, 688, 694
  animation, 456
  CPK models, 450
  direct volume rendering, 453
  geometric representation, 450
  illustration, 455
  ray tracing, 451
  volumetric representation, 452
Molecular machines, 620
Molecular masks, 352
Molecular packing, 148
  efficiency, 709
  measurement of, 709
Molecular replacement, 9
  combined, 349
  data-collection strategies, 227
  electron-density averaging, 352
  in fibre diffraction, 587
  noncrystallographic symmetry, 333
  quality indicators for, 70
  rotation functions, 340
  translation functions, 347
  using an NMR structure, 615
  with *MOLREP*, 364
  with twinning, 548
Molecular surface, 451, 707, 713
Molecular volumes, 703
  Delaunay triangulation, 705
  Voronoi construction, 703
Molecular weight
  measurement using SANS, 579
  measurement using SAXS, 567
Molikin, 688, 692
MolProbity, 689–690, 692, 694
MolProbity score, 697
MolScript, 450, 455
Monochromatic data collection, 211
  data-collection geometries, 212
  detectors, 212
  exposure time, 224
  fine slicing, 215
  rotation range, 215–217, 219
  still exposure, 213
  wide slicing, 215
Monochromators
  crystal, for neutrons, 170
  crystal, for X-rays, 166
  curved single-crystal, 196
  double-crystal, 196
  for synchrotron radiation, 196
  multilayer, for neutrons, 170
  polarizing multilayer, for neutrons, 171
Monoclinic crystal system, 47
Monte Carlo method, in *PrimeX*, 535
Monteath Robertson, J., 5–6
Moore's law, 449
Mosaic spread, 58, 163
Mosaicity, 58, 103, 113, 214, 288
  and data integration, 266
  anisotropic, 300
  ideal, 59
MOSFLM, 265, 269
MRC, 630
MTZ2HKL, 532
Multiconformer models, 472
Multi-copy search, in *MOLREP*, 365
Multidimensional isomorphous replacement (MDIR) in fibre diffraction, 587
Multidimensional NMR, 616
Multidomain averaging, 360
Multifiltering cyclic phasing procedure, 438
Multilayer monochromators for neutrons, 170
Multiple-crystal-form averaging, 360
  in *DM/DMMULTI*, 408, 412
Multiple isomorphous replacement (MIR), 368, 373
  preparation of heavy-atom derivatives, 317
  with twinning, 548
Multiplicity of a data set, 66
Multi-reference alignment, 626
Multistart refinement, 471
Multivariate statistical analysis, 601, 626, 739
Multiwavelength anomalous diffraction (MAD), 373
  advantages of cryocrystallography, 376
  and synchrotron radiation, 199
  anomalous-scatterer labels, 377
  automated structure solution, 379
  conversion of data to a pseudo-SIRAS form, 379
  data-collection strategies, 227
  data handling, 376
  design of experiments, 375
  effect of errors, 376
  phasing, 374, 376
  phasing signal strength, 375
Multiwire proportional counters
  for neutrons, 171
  for X-rays, 180
Myoglobin, 6, 8, 756

$\mathrm{N{-}H\cdots}$carbonyl hydrogen bonding, 741
$\mathrm{N{-}H\cdots}\pi$ hydrogen bonding, 742
Nanobodies, 132
Nanocrystal analysis, 41
Nanomanipulator, 457
Nanotechnology, 41
NDB. *See* Nucleic Acid Database
Neurological disorders, 23
Neutron-beam collimators, 169
Neutron-beam filters, 169
Neutron beamline optics, 173
Neutron density maps, 553
Neutron detectors, 171
  image plates, 172
  multiwire proportional counters, 171
Neutron diffraction, 553
  and solvent structure, 801
  $D_2O - H_2O$ difference maps, 554–555
  geometries, 553
  phasing, 554
  quasi-Laue, 553
  refinement, 555
  time-of-flight, 174
Neutron guides, 171
Neutron instrument resolution functions, 172
Neutron scattering
  coherent, 575
  incoherent, 575
Neutron scattering lengths, 553
Neutron sources, 168
  reactors, 168
  spallation, 172
NEWHELIX, 770
Nicol prism, 147
Nominal resolution, 64–65
Nonbonded interactions
  IsoStar, 743–744
  restraints, 482
Noncrystalline fibres, 583
Noncrystallographic asymmetric unit, 354
Noncrystallographic point-group symmetry, 333
Noncrystallographic redundancy, 354
Noncrystallographic symmetry, 333, 352
  cross-rotation function, 335
  cross-translation function, 335
  determination of, 390
  eigendensity functions, 337
  electron-density averaging, 352, 391
  electron-density averaging in *DM/DMMULTI*, 409, 411
  generalized, 333
  improper, 352–353
  in *Coot*, 445
  in isomorphous replacement, 331
  in phasing, 354
  in *PHENIX*, 540
  in structure determination, 335, 337
  in structure validation, 656, 658
  overdetermination ratio, 336
  proper, 352–353
  refinement of, 391
  restraints in *SHELXL*, 532
  rotation functions, 333
  self-rotation function, 334
  standard, 333

Noncrystallographic symmetry
  subunits, 333
  translation functions, 334
  use of Patterson function, 333
Nonlinear constraints, 396
Non-merohedral twinning, 311
Non-primitive (centred) unit cell, 46
Normal equations, 463, 500
Normal-mode analysis, 474
Normal probability analysis, 68
Normalized structure factors, 57, 415
NUCheck, 833
Nuclear magnetic resonance (NMR), 615
  and single-crystal X-ray diffraction, 617
  and solvent structure, 801
  cis–trans isomerization of peptide bonds, 615
  conformational equilibria, 618
  dynamic processes, 615, 618
  multidimensional, 616
  Nucleic Acid Database, 833
  Protein Data Bank, 827
  proton exchange, 615
  resonance assignments, 616
  ring flipping, 615, 618
  structure determination, 615
  studies of solvation, 617
  transient local conformational states, 615
  triple-resonance experiments, 616
  water-molecule location, 617
  water-molecule residence times, 617
Nuclear Overhauser effect (NOE), 616
  upper-distance constraints, 616
Nucleation, 99, 103, 111, 115
Nucleic Acid Database (NDB), 833
  structure validation, 663
Nucleic acids, 766
  A, B and Z helices, 773
  A-DNA, 766
  B-DNA, 766
  backbone geometry, 766
  base pairing, 767
  crystallization of, 100–102, 115
  DNA, 766
  DNA/RNA hybrids, 766
  duplex RNA, 766
  duplexes, 766
  glycosyl bond geometry, 773–774
  helix parameters, 766, 770
  Hoogsteen base pairing, 769
  hydrogen bonding in, 722, 726
  interactions with proteins, 757
  Nucleic Acid Database, 833
  stacking of base pairs, 776
  structure validation, 662
  sugar ring conformations, 766, 771, 775
  Watson–Crick base pairing, 769
  Z-DNA, 766
Nucleophilicity, 318–319
N-terminal heterogeneity, 88

O, 464
O—H···O hydrogen bonding, 741
O—H···π hydrogen bonding, 742
Obliquity correction for CCD detectors, 186
Observational equations, 462
Obsolete PDB entries, archive of, 680, 682
Occam's principle, 280
Oligonucleotides, 766
OmpA, 124
OmpF, 124
OOPS, 651
Opening angle of synchrotron radiation, 190
OPLS force field, 535–537
Optical properties of crystals, 145–146
Optical resolution, 65
Optimization of crystallization conditions, 105–107, 110–111, 113, 115
Optimization methods, 460, 462
  first-order, 462
  Monte Carlo, 462
  second-order, 462
  simulated annealing, 462
  zero-order, 462
Order–disorder phenomena, 315
Origin-removed Patterson refinement, 379
ORTEP, 449, 455
Orthogonal molecular interactions in halogen bonds, 824
Orthogonalization matrix, 355
Orthorhombic crystal system, 47
Outliers
  in data integration, 267, 270
  in derivation of restraints, 477
  in protein-structure models, 464, 650
Overdetermination ratio, 336
Overfitting, 466, 651
Overlap of electron density in noncrystallographic-symmetry-related regions, 70
Overloads, 271
Oversampling phasing method, 234

p-cell, 354
Packing coefficient, 709
Packing density, 709
Packing efficiency, 709
Papain, 9
Parallel-axis theorem, 578
Parallel processing, 309
  in *EMAN*, 630
Parameter-shift method, 418
Parseval's theorem, 401
Partial occupancy, 485, 490
  water-molecule sites, 493
Partial specific volume, 152
Partial spots, 273
Partial twinning, 312
Partiality, 289, 296
Partiality model, 301
Partially recorded reflections, 214
  in data processing, 296
  profile fitting, 271
Particle mesh Ewald method, 643
Patterson, A. L., 6
Patterson-correlation translation function, 348
Patterson functions, 62, 328
  and noncrystallographic symmetry, 333
  cross vectors, 328–329
  cylindrically averaged, 586
  modified, 329
  origin peak, 328
  self vectors, 329
  use of, in *Shake-and-Bake*, 419
Patterson maps, 62
  anomalous difference, 331
  Bijvoet, 9
  diffraction ripples in, 62
  number of peaks in, 62
  sharpening of, 62
Patterson minimum function, 330
Patterson synthesis, 6
Pauling, L., 6
PCR (polymerase chain reaction), 76–77
PDB Exchange Dictionary (PDBx), 827
PDB Japan (PDBj), 830
PDBeFold, 830
PDBeMotif, 830
PDBePISA, 830
PDBeXplore, 831
Peak picking, 418
Peaklist optimization, 418
Pearson correlation, 550
PEGs (polyethylene glycols), 86, 100–101, 123, 249
Penicillin, 5
Pepinsky, R., 6, 448
Pepsin, 6
Peptide flip, 650, 655
Peptides
  cis, 654
  cis–trans isomerization, NMR studies of, 615
  trans, 654
Perfect twinning, 312
Perturbation-γ correction, 395
  in *DM/DMMULTI*, 409
Perutz, M. F., 6
pET vectors, 92
Phase averaging, 440
Phase circle, 549
Phase combination, 385, 394, 404
  in *DM/DMMULTI*, 409
Phase contrast, 595, 599
Phase determination, quality indicators for, 69
Phase extension
  and noncrystallographic symmetry, 359
  by electron-density averaging, 352
  in *DM/DMMULTI*, 409
  in reciprocal space, 417
Phase improvement, 385
  constraints, 396
  in *DM/DMMULTI*, 407
  quality indicators for, 69
Phase probability, 330
Phase probability distributions, 369
  Bayesian calculation of, 379
  for anomalous scattering, 370
Phase problem, 40, 327
Phase refinement
  by electron-density averaging, 352
  in reciprocal space, 417
Phase retrieval, from single-particle diffraction data, 235
Phase retrieval transfer function, 235
Phase sets
  alignment of, 440
  cluster analysis of, 441
  comparison of, 439
  generation of, 438
Phased translation function, 347, 349
PHASES, 361
Phasing
  ab initio, 413
  ab initio, in molecular replacement, 360
  in neutron diffraction, 554
  in *PHENIX*, 540
  MAD, 376
  multisolution methods, 417
  power of, 360
  using anomalous scattering, 367
  using noncrystallographic symmetry, 354
Phasing figure of merit, 330, 369, 381
Phasing power, 69, 330
  anomalous, 69
PHENIX, 539, 701
  graphical user interface, 539
Phenotypic selection methods, 130
PhoE, 124
Phong shading, 451
Phosphate-binding protein, 761
  short hydrogen bond to phosphate, 763
Photographic film, 594, 599–600
Photon-counting X-ray detectors, 179
Photon-integrating X-ray detectors, 179–180
Photosynthetic reaction centres, 123
Planarity restraints, 481
Plasma X-ray sources, 159
Pluto, 455
Point groups, 47
  centrosymmetric, 50
  enantiomorphic, 47–48
  icosahedral, 52
Point inducible dipole, 730
Point X-ray detectors, 179
Poisson–Boltzmann equation, 731
Polarity, 731
Polarizability, 730–731
  hard and soft metals, 756
Polarization factor, 59–60
Polarization of electron density, 823
Polarization of synchrotron radiation, 189
Polarizing multilayer monochromators for neutrons, 171
Polycrystalline fibres, 583
Polyethylene glycols (PEGs), 86, 100–101, 123, 249
Polymerase chain reaction (PCR), 76–77
Polymorphism, 146, 333
Polymorphs, 113, 115
Polynucleotide chain tracing, 527
Pooled coefficient of variation, 67, 653
Porins, 124
Position error, 507
Positional search, in *MOLREP*, 365
Positional shifts, 509

Post refinement, partially recorded reflections, 296
Post-translational modifications, 75–76, 82, 89
  removal of, 133
Potassium channel, 123
Potential-energy functions, 642
Power of phase determination, 360
Precipitants, 123, 319
  ammonium sulfate, 86, 319
  magnesium sulfate, 319
  sodium/potassium phosphate, 319
Precision, 499
  full-matrix estimates of, 505
  relative, 665
Preconditioned conjugate-gradient method, 463, 523
  in *TNT*, 522
Prenucleation, 111
Prime, 534, 537
PrimeX, 534
Primitive unit cell, 46
Principal component analysis, 739
Probe, 688, 690–692, 694–695, 698–699
Probe radius, 709
Probe sphere, 707
PROCHECK, 464, 662–663, 677, 684
PROCHECK-NMR, 684
Profile fitting, 266, 268
  in fibre diffraction, 586
  partially recorded reflections, 271
  strong reflections, 270
  systematic errors, 271
  weak reflections, 270
Project MAC, 449
cis-Prolines, frequency of, 474
PROLS, 463
PROSA II, 664
Prostaglandin H_2 synthase, 124
PROTEIN, 330
Protein–carbohydrate recognition, 755
Protein chain tracing, 526
Protein crystallization, 129
Protein Data Bank (PDB), 827
  deposition sites, 827
  IsoStar, 743–744
Protein Data Bank in Europe (PDBe), 830
Protein–DNA recognition, water molecules in, 816
Protein domains, 752
  boundaries, 752
  discontinuous, 752
  identification, 749, 752
Protein engineering, 75, 129, 140
Protein expression, 75–77, 142
  baculoviruses, 83
  constitutive, 78
  constructs, 76
  fermentation, 80
  growth media, 80
  in *E. coli*, 77, 79, 142
  in yeasts, 82
  inducible, 78
  insect cell–virus, 83
  mammalian cells, 84
  misfolded proteins, 80
  plasmids, 78
  post-translational modifications, 75–76, 82, 89
  preparation of cDNA clones, 76
  T7 polymerase, 78
Protein families, 749
Protein folding, 81
  hydrogen bonding in, 723
  in vivo, 81
  misfolded proteins, 80
  refolding, 87
  role of water molecules, 800
  studied by SAXS, 572
  use of chaperones, 81
Protein function
  role of metal ions, 756
  role of water molecules, 800
Protein heterogeneity, conformational, 104
Protein kinase A, solvent structure, 814
Protein–ligand docking programs, 745
Protein–ligand interactions, 755
  carbohydrates, 755
  IsoStar, 743
Protein–ligand interactions
  metals, 756
  phosphate, 761
  role of water molecules, 800, 807
  sulfate, 761
  use of the CSD, 740, 743
Protein–nucleic acid interactions, 757
  DNA, 757, 759
  RNA, 759–760
  transfer RNA, 759
Protein purification, 85
  chromatography, 86
  isoelectric focusing, 88
  mass spectrometry, 88
  N-terminal heterogeneity, 88
  sample heterogeneity, 89
  SDS–PAGE, 88
Protein refolding, 87
Protein stability
  effect of hydrogen bonding, 723
  role of metal ions, 756
Protein structure classification, 749, 752
  CATH, 749–751
  DALI, 749–751
  DALI domain dictionary, 751, 753
  3Dee, 749–750
  ENTREZ, 749, 751
  HOMSTRAD, 749–750
  SCOP, 749–750
  SSAP, 749
  STAMP, 749
  VAST, 751
Protein superfamilies, 750
Protein targets, stabilization of, 134
Protein–water interactions
  database analysis, 802
  effect of secondary structure, 805
  effect of tertiary structure, 806
ProteinCCD server, 136
Proteins, storage of, 89
Proteolysis, 131
  in *E. coli*, 81
  N-end rule, 79
  of recombinant proteins, 81
Proton exchange, NMR studies of, 615
Protozoan infections, 16, 21
PROVE, 664
PS79, 587
Pseudo-crystallographic symmetry, 333
Pseudo-merohedral twinning, 312
Pseudosymmetry, 654
  in isomorphous replacement, 331
Pump–probe studies, 42
Pycnometry, 154
Python, 630

Q factor, 601
Quadratic information measure, 487
Quality indicators, 650, 652
  G factors, 656
  merging *R* factors, 653
  pooled coefficient of variation, 653
Quality of data, 65
Quasi-Laue neutron diffraction, 553

R factors, 464
  crystallographic, 71, 651, 657, 665
  for density modification, 70
  in fibre diffraction, 588
  R_{anom}, 67
  R_{Cullis}, 69, 330
  $R_{\mathrm{Cullis,ano}}$, 69
  R_d, 68
  R_{free}, 70, 71, 464, 466, 506, 651, 657, 665
  R_{Kraut}, 330
  R_{meas}, 66
  R_{merge}, 66, 224, 228, 298, 559
  $R_{\mathrm{mrgd}\text{-}F}$, 66
  $R_{\mathrm{mrgd}\text{-}I}$, 66
  $R_{\mathrm{p.i.m.}}$, 66
  $R_{\mathrm{r.i.m.}}$, 66
  R_{scale}, 68
  R_{sym}, 559
  real-space, 657, 666
R-factor translation function, 347
RADDOSE, 258
Radiation damage, 68, 225, 249, 256
  and cryogenic vitrification, 225
  and Laue diffraction, 207
  free radicals, 243
  in electron diffraction, 558
  in electron microscopy, 593–594
  in single-particle diffraction, 236
  suppression of, 242
Radiation-damage-induced phasing (RIP), 260
Radiation dose, 256
Radiation dose limit, 258
Radiation-induced shrinkage, 236
Radicals, 259
Radioactive X-ray sources, 159
Radius of convergence, 459
Radius of gyration, 565, 567, 576
Ramachandran plot, 454, 650, 654, 678, 685, 739
  analysis, 695
  multiple-model, 656
Ramachandran restraints, in *Coot*, 444
Random-atom model, 435
Random-conical reconstruction, 620
Random omit maps, 419
RasMol, 450
Raster3D, 717
RAVE, 361
Ray tracing, 451
Rayleigh scattering, 59
Reaction potential, 731
Real-space calculations, in *PrimeX*, 537
Real-space constraints, 418
Real-space fit, 650, 657
Real-space *R* factor, 657, 666
Real-space refinement, in *Coot*, 444
Real-space residual, 71
Real-space search, in *MOLREP*, 366
Recentring, 435
Reciprocal lattice, 57
Reciprocal-lattice points, non-integral, 354
Reciprocal-lattice vectors, distribution of, 263
Reciprocal space, 57
Reciprocal-space calculations, in *PrimeX*, 536
Recombinant proteins, 75
  incorporation of selenomethionine, 76
  minimizing proteolysis of, 81
  toxicity of, to host, 81
Recombination cloning, 141
Reconstruction engine, 620
Reduced cell, 264, 279
Reduced Miller indices, 296
Redundancy, 66, 206–207, 217, 227, 376, 653
  noncrystallographic, 354
Re-entrant surface, 707, 713
Reference bias, 626
Refinement, 459
  against intensities, 460
  atomic resolution, 485
  bias, 405
  block-matrix approximation, 488
  conjugate-gradient method, 397, 463, 488
  conjugate-gradient method, preconditioned, 463, 522
  coordinate uncertainty, 499
  data quality, 460
  diagonal-approximation method, 488
  difference-Fourier method, 459
  global minimum, 466
  in electron diffraction, 560
  in fibre diffraction, 587
  in *SHELX*, 529
  in *SHELXL*, 531
  intensity-based likelihood, 489
  least-squares, 459–460, 488, 500
  local minima, 466
  maximum-likelihood, 460, 467, 489
  modelling of solvent, 462, 464
  models, 461
  neutron diffraction data, 555
  normal equations, 463, 500
  observational equations, 462

Refinement
  partial occupancy, 490
  quality indicators for, 71
  radius of convergence, 459
  restrained, 501, 504
  restraints, 474–475
  rigid groups, 461
  simulated annealing, 463, 466
  singularity in, 463
  target functions, 466–467
  torsion-angle, 466
  weighting, 460, 500
  with twinning, 549
Reflection intensity, integrated, 58
Reflection-omit method, 395
  in *DM/DMMULTI*, 409
Reflection profiles, 223
REFMAC, 489
Refraction, 146
Refractive index, 146, 148
Register error, 649
Remote data collection, 231
Representation of information, 454
Representation of structures, 448–449
  animation, 456
  CPK models, 450
  direct volume rendering, 453
  geometric, 450
  illustration, 455
  isosurfaces, 452
  physical models, 456
  ray tracing, 451
  stereolithography, 456
  three-dimensional printing, 456
  volumetric, 452
Representation of surfaces, 450, 713, 717
  colour coding, 451
  Gouraud shading, 451
  Phong shading, 451
  photorealistic rendering, 717
  roadmaps, 719
  shaded backbone, 717
  texture mapping, 451
Research Collaboratory for Structural Bioinformatics (RCSB PDB), 829
Residual function, 501
Resolution, 243
  effective, 66
  in single-particle recontruction, 632
  nominal, 64–65
  optical, 65
  true, 65
Resonance-assisted hydrogen bonding, 741
Resonance-induced hydrogen bonding, 741
RESTRAIN, 488
Restrained full-matrix inversion for concanavalin A, 502
Restrained refinement, 501, 504
  in fibre diffraction, 587
  residual function, 501
  two-atom model, 501
Restraints, 71, 461, 474–475, 535
  atomic displacement parameter, 490
  bond-angle, 474, 477, 479
  bond-length, 474, 476–477
  choice of, 475
  coordinate, 490
  effect on error estimates, 488
  ensemble-averaged, 633
  geometrical, 461
  in molecular dynamics, 633, 643
  in *PHENIX*, 543
  in *SHELXL*, 531
  in *TNT*, 520
  nonbonded interactions, 482
  planarity, 481
  special geometries, 482
  target parameters, 474
  time-averaged, 633
  torsion-angle, 482
  use of the CSD, 474
  weighting of, 502
Reticular twinning, 312
Rhodopsin tag, 93
Ribonuclease, 9
  A, solvent structure, 813
  T1, solvent structure, 813
Ribose-pucker outliers, 697
Ribosome, synchrotron-radiation studies of, 198
Ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO), structure validation, 679
Richards box, 10, 449
Rigid-body superposition, 749
Rigid-group refinement, 461
  in *TNT*, 522
Ring flipping, NMR studies of, 615, 618
RIP (radiation-damage-induced phasing), 260
RNA
  duplex RNA, 766
  hydrogen bonding in, 726
  interactions with proteins, 759–760
  Nucleic Acid Database, 833
  treatment in *PHENIX*, 541
RNA backbone suite outliers, 699
RNA dimerization initiation site, 825
RNABC, 699
Robertson sorting board, 6
Robotic crystal loading, 231
Robotics, in crystallization, 107
Robots, 247
Rocking curve, 214
Rocking width, 197
Root-mean-square correlation ratio, 68
Root-mean-square deviation from crystallographic symmetry, 73
Root-mean-square Z score, 73
Rossmann, M. G., 9
Rotamer conformations, 655
Rotamer side-chain fit, 651
Rotating-anode X-ray tubes, 159
Rotation axes, 46–47
Rotation functions, 340
  and noncrystallographic symmetry, 333
  computation of, 342
  cross-rotation, 341
  fast, 342
  locked, 343
  quality indicators for, 70
  sampling of, 342
  self-rotation, 341
  symmetry properties of, 343
Rotation group, 342
  metric of, 341
Rotation method, 213, 272
Rotation range, 215–217, 219
Rotational disorder in fibre diffraction, 584
Rotational search, in *MOLREP*, 365
Rotational symmetry, 46
Rotations in three-dimensional Euclidean space, 340
Royal Institution, The, 845
Rusticyanin, structure validation, 667
RvR calculation, 549

σ holes, 823, 825
σ_A plot, 73, 404, 658
σ_A values, estimation of, 403
σ_A weighting, 394
Saddlepoint method, 435
Sample acceptance, 191
SANS. *See* Small-angle neutron scattering
SAXS. *See* Small-angle X-ray scattering
Sayre's equation
  application to macromolecules at non-atomic resolution, 392
  for phase refinement and extension, 392
  in real and reciprocal space, 392
  shape function, 392
Scale factor, 286, 290–291
Scaling, 68, 277
  anisotropic, 488
  Hamilton, Rollett and Sparks method, 296
  in density modification, 390
  in *DM/DMMULTI*, 410
  in *MOLREP*, 365
  in *SFCHECK*, 666
  local, 381
  of intensity data for lysozyme, 851
  of structure factors, 57
Scaling
  partially recorded reflections, 296
  R factors, 68, 298
  restraints and constraints, 297
  selection of reflections for, 297
SCALEPACK, 282
Scattering, 59
  Compton, 59
  elastic, 52
  electron, 557
  inelastic, 52
  Rayleigh, 59
  Thomson, 53
Scattering factors
  anomalous, 373
  atomic, 54, 373
  atomic, solvent-corrected, 587
Scattering lengths, neutron, 553
Scavengers, 259
SCOP, 749–750
Screening potential, 731
Screenless rotation method, 212
Screw axes, 46, 61
Screw disorder in fibre diffraction, 584
SDS–PAGE, 88
Search model, in *MOLREP*, 364
SEC (size-exclusion chromatography), 95
Secondary structure
  and hydrogen bonding, 723, 742
  effect on protein–water interactions, 805
  from NMR studies, 615
  of RNA, 759
Selection rule in fibre diffraction, 584
Selenocysteine, 325
Selenomethionine, 76, 317, 325, 377
Self-potential, 731
Self-rotation function, 334, 341
Self vectors, 329
Sequence tags, 77
Serine proteases, solvent structure, 808
SFCHECK, 657, 666, 834
  global quality indicators, 666, 668
  local quality indicators, 667, 670
Shake-and-Bake, 423
'Shaking', 417
Shannon's sampling theorem, 358
Shape function, 392
β-Sheets, 6
  hydrogen bonding in, 724
SHELX, 529
SHELXL, 487, 665
SHELXL98, 463–464
SHELXPRO, 532
Shine–Dalgarno sequence, 78–79
Shutterless detector systems, 41
Side-chain hydrogen bonding, 724
Side-chain placement, in *PrimeX*, 537
Side-chain rotamers, 695
Side-chain torsion angles, 655
SIFTS, 831
SIGMAA, 403, 405
Signal sequence, 93
Signal strength and structure validation, 653
Signal-to-noise ratio, 64
  highest possible, 67
  mean, 67
  mean anomalous, 68
Sim distribution, 402, 467
Sim weighting, 394
  in *DM/DMMULTI*, 409
Simple point charge (SPC) model, 633–634
Simulated annealing, 463, 466
  annealing schedules, 470
  comparison with conjugate-gradient method, 471
  in *CNS*, 517
  in fibre diffraction, 588
  in *PrimeX*, 536–537
  molecular dynamics, 469
  multistart refinement, 471
  searching conformational space, 468
  temperature, 469
Single isomorphous replacement (SIR), 371
Single-particle diffraction, 42, 235
Single-particle electron cryomicroscopy, 42, 624

Single-particle imaging, 41
Single-particle reconstruction, 620, 629
  resolution of, 43
Single-wavelength anomalous scattering (dispersion), data-collection strategies, 227
Singularity in refinement, 463
Site-directed mutagenesis, 130, 317, 325
Size-exclusion chromatography (SEC), 86, 95
Skeletonization, 392
  in *DM/DMMULTI*, 407, 409
Skewness of electron density, 70
Small amphiphile concept, 125
Small-angle neutron scattering (SANS), 575
  crystallization studies, 111
  Debye equation, 575
  distance measurements, 579–580
  length distribution, 576
  molecular weights, 579
Small-angle X-ray scattering (SAXS), 563
  computer programs, 571
  crystallization studies, 111
  data collection, 570
  data processing, 571
  experiment design, 570
  instrumentation for conventional sources, 568
  low-resolution model determination, 567
  sample handling, 569
  sample preparation, 569
  single-crystal, 563
  solution, 564
  synchrotron instrumentation, 568
SnB, 421
Sodium/potassium phosphate, 319
Soft ligands, 318
Soft metals, 757
Solubility of proteins, 99, 101, 114–115, 130
  solubility enhancers, 103
  solubilizing agents, 103
Solution X-ray scattering, 564
  and crystal structures, 567
  computer programs, 571
  data collection, 570
  data processing, 571
  Debye formula, 567
  Debye function, 565
  electron pair distribution function, 565–566
  experiment design, 570
  Guinier plot, 565
  instrumentation for conventional sources, 568
  low-resolution model determination, 567
  molecular weights, 567
  radius of gyration, 565, 567
  sample handling, 569
  sample preparation, 569
  synchrotron instrumentation, 568
Solvation, NMR studies of, 617
Solvation free energies, 822
Solvation potential, 731
Solve, 330, 379, 382
  scoring trial heavy-atom solutions, 380
Solvent
  bound, 153
  bulk, modelling of, 462, 464, 493
  in structure validation, 656
  modelling in *SHELXL*, 532
  ordered, at atomic resolution, 493
  variation of density, 577
Solvent-accessible surface, 451, 707, 713
Solvent-corrected atomic scattering factor, 587
Solvent density, 156
Solvent-excluding surface, 451, 713
Solvent flattening, 386
  in *DM/DMMULTI*, 407–408
  in fibre diffraction, 587
Solvent flipping, 388, 395
Solvent masks, in *DM/DMMULTI*, 407, 411
Solvent modification, 241
Solvent structure, 800
  $D_2O - H_2O$ difference maps, 555
  neutron diffraction, 554, 801
  NMR studies, 801
  simulation of, 801
  solution X-ray scattering, 567
Source entropy, 433
Space-group assignment, 278, 654
Space groups, 46
  common, for protein crystals, 47
Spallation neutron sources, 172
Special positions, 47
Spectral brightness of synchrotron radiation, 189
Sphere of influence algorithm, in *SHELXE*, 530
Sphere refinement, in *Coot*, 444
Spherical aberration, 599
Spherical Bessel functions, 345
Spherical harmonics, 345
SPIDER, 620, 630
Spin contrast variation, 578, 580
SPIRE, 620
Squalene cyclase, 124
SQUASH, 361
SSAP, 749
STAMP, 749
Standard atomic radii, 708
Standard atomic volumes, 710
  in structure validation, 664
Standard basis vectors, 46
Standard profiles, 268
Standard residue volumes, 710
Standard spot shape, 273
Standard uncertainty (s.u.), 65, 499, 501, 665
  estimation in *SHELXL*, 531
Stanley, W., 7
Stanley factor, 313
Staphylococcal nuclease, 757
Static disorder, 462, 485
Stationary-target X-ray tubes, 159
Stereolithography, 456
Stereoscopic viewing, 237
Still exposure, 213
Storage of crystals at low temperature, 253
Storage of proteins, 89
Structural genomics, 13, 199
Structure-based drug design, 15
  use of the CSD, 745
Structure determination
  by fibre diffraction, 586
  by NMR, 615
  by single-crystal X-ray diffraction and NMR, 617
  in *CNS*, 512
  in *SHELX*, 529
  using noncrystallographic symmetry, 335, 337
Structure-factor averaging, 471
Structure-factor probability distributions, 401
  general treatment, 403
  Luzzati, 403
  Sim, 402, 467
  Wilson, 402
  Woolfson, 402
Structure factors, 55–56
  centrosymmetric structures, 56
  Fourier–Bessel, 584
  in electron diffraction, 559
  in fibre diffraction, 584
  noncentrosymmetric structures, 56
  normalized, 57, 415
  placing on an absolute scale, 57
Structure invariants, 416
Structure refinement, in *PHENIX*, 543
Structure representation, 448–449
  animation, 456
  CPK models, 450
  direct volume rendering, 453
  geometric, 450
  illustration, 455
  isosurfaces, 452
  physical models, 456
  ray tracing, 451
  stereolithography, 456
  three-dimensional printing, 456
  volumetric, 452
Structure solution
  automated, for MAD and MIR, 379
  in *PHENIX*, 539
Structure validation, 649, 662, 677
  agreement of model with experimental data, 662, 665
  at the NDB, 663
  Balasubramanian plot, 655
Structure validation
  bond angles, 654, 662, 685
  bond lengths, 654, 662, 685
  C^{α}-only models, 655
  choice of reference structures, 679
  data completeness, 653
  data resolution, 653
  detection of outliers, 650
  difference density quality, 658
  diphtheria toxin, 681
  directional atomic contact analysis, 655
  environment profiles, 663
  geometric parameters, 654, 662
  hetero groups, 663
  hydrogen-bonding analysis, 651, 655
  in *MolProbity*, 694
  knowledge-based interaction potentials, 663
  main-chain torsion angles, 677
  nonbonded interactions, 677
  nonbonded parameters, 663
  noncrystallographic symmetry, 656, 658
  nucleic acids, 662
  packing, 677
  pep-flip value, 650
  planarity, 654
  PROCHECK, 684
  quality indicators, 650
  Ramachandran plot, 650, 654, 678, 685
  real-space fit, 650, 657
  rotamer side-chain fit, 651
  RuBisCO, 679
  rusticyanin, 667
  SFCHECK, 666
  side-chain torsion angles, 655
  signal strength, 653
  solvent, 656
  standard atomic volumes, 664
  standard values, 662–663
  stereochemical parameters, 654, 662–663
  torsion angles, 654
  triacylglycerol lipase, 681
  unit-cell parameters, 653
  use of the CSD, 662, 737
Stuhrmann equation, 577
Substituent effects in halogen interactions, 821
Substructure determination, 68
  in *PHENIX*, 540
Substructure solution, with *SHELXD*, 529
Subtilisin, 9
Sugar ring conformations, 766, 771, 775
  C2′-*endo*, 767
  C3′-*endo*, 767
  envelope (E), 766
Sulfate-binding protein, 761
Summation integration, 266–267
Supermirrors for neutrons, 170
Supersaturation, 99
SuperStar, 745
Support film, 597, 599, 603–604
Supramolecular synthons, 743
Surface-area calculation, 713–714
  analytical, 715
  complete rolling algorithm, 714
  connected rolling algorithm, 714
  Connolly dot surface algorithm, 714
  extended atoms, 715
  Gauss–Bonnet theorem, 715
  Lee & Richards planar slices, 714
  marching-cube algorithm, 714
Surface areas, 703
Surface-entropy reduction, 134
Surface-entropy reduction server, 136
Surfaces, 713
  Connolly surface, 451, 717
  contact surface, 707, 713
  convex hull, 706
  definitions of, 706
  GRASP surfaces, 718
  hydration surface, 714
  molecular surface, 451, 707, 713
  occluded molecular surface, 713
  probe sphere, 707
  re-entrant surface, 707, 713
  solid polyhedral surfaces, 717

Surfaces
  solvent-accessible surface, 451, 707, 713
  solvent-excluding surface, 451, 713
  van der Waals surface, 707, 713
  Voronoi polyhedra, 706
Surfaces, representation of, 450, 713, 717
  colour coding, 451
  Gouraud shading, 451
  Phong shading, 451
  photorealistic rendering, 717
  roadmaps, 719
  shaded backbone, 717
  texture mapping, 451
SurVol, 664
Symmetry, 46
  helical, 333, 584
  icosahedral, 47
  in diffraction patterns, 60
  inversion, 46
  mirror, 46
  missing, 654
  rotational, 46
  translational, 46
Symmetry correction, 336–337
Synchrotron radiation, 189, 212
  and Laue diffraction, 195, 205
  and MAD, 199
  and SAXS, 568
  angular brightness, 189
  area detectors, 195, 198
  atomic resolution studies, 198
  beamlines for macromolecular crystallography, 192
  bending magnets, 161
  collimation, 189
  critical wavelength, 161, 190
  detectors, 195
  F_1 ATPase, 198
  flux, 189
  in protein crystallography, 197
  insertion devices, 161, 190–191
  instrumentation, 189, 194
  intensity, 192
  machine emittance, 190
  monochromatic, 196
  monochromators, 196
  multipole wigglers, 190
  opening angle, 190
  polarization, 189
  production of, 189
  properties of, 189
  sample acceptance, 191
  sources, 160
  sources, first-generation, 192
  sources, second-generation, 193
  sources, third-generation, 193
  spectral brightness, 189
  studies of multi-macromolecular complexes, 198
  studies of small crystals, 198
  studies of the ribosome, 198
  time-resolved studies, 198
  total radiated power, 190
  tunability, 189
  undulators, 161, 190
  universal curve, 190
  wavelength shifters, 190
  wigglers, 161
Synthetic biology, 824
Systematic absences, 61
π systems in halogen bonds, 824–825

T4 lysozyme, solvent structure, 812
T7 polymerase expression system, 78
Tags, 77, 131
  removal of, 77
Tangent formula, 417
Target constructs, optimization of, 131
Target functions, 466–467
  in *CNS*, 512, 514
  MLF, 468
  MLHL, 468
  MLI, 468
Target parameters, 474
Telluromethionine, 317
Temperature (in simulated annealing), 469
Temperature factors (atomic displacement parameters), 56–57, 485
  anisotropic, 56, 485
  anisotropic, at atomic resolution, 492
  anisotropic, refinement in *SHELXL*, 531
  constraints, 490
  effect of coordinate errors, 461
  effect on coordinate uncertainty, 499
  group *B* factors, 461
  in structure validation, 656
  isotropic, 56
  refinement of, 461
  restraints, 490, 522
Tertiary structure
  effect on protein–water interactions, 806
  of RNA, 759
Tetartohedral twinning, 315
Tetragonal crystal system, 52
Texture mapping, 451
Thermostability, 134
Thomson scattering, 53
Thon rings, 42
Three-dimensional image reconstruction, 596–597, 601–602
Three-dimensional printing, 456
Three-dimensional reconstruction, 42, 626
  in electron diffraction, 558
  software, 43
Thyroxine, 821
Time-resolved crystallography, 41
Time-resolved imaging, 42
Time-resolved studies, 198, 205, 208
  in fibre diffraction, 585
  using SAXS, 572
TNT, 463
Tobacco mosaic virus (TMV), 6
  fibre diffraction, 587
Tomographic crystal-volume measurement, 155
Torsion-angle molecular dynamics, 469–470
Torsion-angle refinement, 466
Torsion-angle restraints, 482
Torsion angles
  in DNA, 771
  in nucleic acid backbone chains, 766
Transfection, 84
Transfer RNA
  interactions with proteins, 759
  low-abundance, 79
Translation functions, 347
  and noncrystallographic symmetry, 334
  correlation-coefficient, 347
  locked, 350
  packing analysis, 349
  Patterson-correlation, 348
  phased, 347, 349
  quality indicators for, 70
  R-factor, 347
Translation, libration and screw tensor, 462, 485, 488, 492
Translational cross-correlation, 621
Translational disorder in fibre diffraction, 584
Translational symmetry, 46
Transmission factor, 59
Transverse relaxation-optimized spectroscopy (TROSY), 616
Triacylglycerol lipase, structure validation, 681
Triangulation number, 47
Triclinic crystal system, 47
Trigonal crystal system, 52
Triple isotopic substitution, 578
Triplet invariants, 416
Tris buffer, 319
True resolution, 65
Trueblood, K. N., 5
Truncation, 131
Trypsin, molecular-dynamics simulation, 645
Tubulin, electron-diffraction studies, 560–561
Tunability of synchrotron radiation, 189
Turns, hydrogen bonding in, 724
Twin fraction, 312, 548
Twinning, 146
  by hemihedry, 312
  by merohedry, 311
  by non-merohedry, 311
  by pseudo-merohedry, 312
  by tetartohedry, 315
  epitaxial, 311
  hemihedral, 146
  in autoindexing, 284
  indications of, 312
  partial, 312
  perfect, 312
  reticular, 312
  tests for, 313
  treatment in *SHELXL*, 532
Two-dimensional crystals, 558, 561, 603

Ubiquitin, molecular-dynamics simulation, 633
Undulators, 161, 190
Uneven crystal rotation, 290
Uneven exposure, 290
Uniaxial crystals, 147
Unified-atom parameters, 708
Unit cell, 45
  non-primitive (centred), 46
  origin choice, 46
  primitive, 46
Unit-cell parameters, accuracy of, 266, 653
United-atom approach, 633
Universal curve, 190
Unrestrained full-matrix inversion
  for an immunoglobulin, 504
  for concanavalin A, 502

Vaccines, 26
Validation
  in *Coot*, 445
  in *PHENIX*, 542
  Protein Data Bank, 827
van der Waals radii, 708, 713, 740, 821–822
  from the CSD, 709
van der Waals surface, 707
Vapour-diffusion methods of crystallization, 104, 106
Variable virtual bond method, 587
Variances, 499
VAST, 751
Velocity selectors, 171
VERIFY3D, 678
Vertex error, 705
Virtual reality, 449
Viruses, 16–17
  crystallization of, 108, 110, 112, 115
  early studies, 9
  helical, 589
  icosahedral symmetry, 47
Visualization, 40
  of information, 454
Vitamin B_{12}, 5
Vitrification of specimens for cryoTEM, 597, 599
Volume *Z* score, 664
Volumenometry, 154
Voronoi construction, 664, 703
Voronoi polyhedra, 703
  and surfaces, 706
  chopping-down method, 705
  for proteins, 704
  method B, 704
  radical-plane method, 705
  ratio method, 704
  vertex error, 705

Water molecules
  as mediators in complex formation, 815
  automatic location of, 493
  buried, 726, 810
  channel sites, 810
  criteria for placing in electron-density maps, 801
  crystal-contact sites, 810
  $D_2O - H_2O$ difference maps, 555
  determining position of, 801
  distribution around amino acids, 802
  in antigen–antibody association, 815
  in dimeric haemoglobin, 816
  in protein–DNA recognition, 816
  interactions with proteins, 802
  location by NMR, 617
  NMR studies, 801

Water molecules
  orientation of, from neutron diffraction, 554–555
  partially occupied sites, 493
  residence times, 617
  role in protein folding, 800
  role in protein structure and function, 800
  surface sites, 810
Watson, J. D., 6
Watson–Crick base pairing, 769
Watson–Crick Z-DNA, 778
Wavelength bandpass, 214
Wavelength normalization curve, 195, 206
Wavelength shifters, 190
Wavevector, 272
Weak intensity data, inclusion in refinement, 460
Weak intermolecular interactions, 742
Weak-phase-object approximation, 599
Weighting, 500
  in *MOLREP*, 365
  of data in refinement, 460
  of diffraction data and restraints, 502
  σ_A, 394
  Sim, 394
WHAT IF, 464, 662–663, 678
Wide slicing, 215
Wiener filtering, 621
Wigglers, 161
  multipole, 190
Wilson distribution, 402
Wilson plot, 57
  B factor, 68
Woolfson distribution, 402

XDS, 272, 304
XDSCONV, 309
X-PLOR, 463, 644
  for fibre diffraction, 588
  restraints, 662
X-RAC, 448
X-ray detectors, 177, 183
  accuracy, 177
  and synchrotron radiation, 195
  area, 179
  dynamic range, 178
  film, 180
  flat-field corrections, 178
  gas discharge (wire) counters, 180
  geometric distortion, 178
  image plates, 180
  line spread function, 177
  multiwire proportional counters, 180
  photon counters, 179
  photon integrators, 179–180
  pixel array, 181
  point, 179
  point spread function, 177
  reverse-biased semiconductor, 180
  scintillator/photomultiplier, 179
  signal-to-noise ratio, 177
  spatial resolution, 177
  stopping power, 178
  storage phosphors, 180
  television, 181
X-ray free-electron lasers, 41, 194
X-ray generation, 159
X-ray mirrors, 164
X-ray sources, 39, 159
  channelling radiation, 159
  for monochromatic data collection, 211
  microfocus X-ray tubes, 160
  plasmas, 159
  radioactive sources, 159
  rotating-anode X-ray tubes, 159
  stationary-target X-ray tubes, 159
  X-ray free-electron lasers, 41, 194
X-ray topography, crystallization studies, 113
X-ray tubes
  characteristic radiation, 160
  microfocus, 160
  rotating-anode, 159
  stationary-target, 159
X-ray wavelength, choice of, 162, 222
XSCALE, 277, 308
XtalPred server, 136
XtalView, 330, 464

Yeast, high-resolution imaging of, 235
Yeasts as expression systems, 82

z-buffer, 450
Z score, 71
  root-mean-square, 73
Z-DNA, 766
Zingers, 185